Solving Matrix Sensing to Optimality under Realistic Settings

Ziye Ma

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-18
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-18.html

April 24, 2024

Solving Matrix Sensing to Optimality under Realistic Settings

by

Ziye Ma

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Somayeh Sojoudi, Chair
Professor Javad Lavaei
Professor Kameshwar Poolla

Spring 2024

Solving Matrix Sensing to Optimality under Realistic Settings

by

Ziye Ma

Abstract

Solving Matrix Sensing to Optimality under Realistic Settings

by

Ziye Ma
Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley
Professor Somayeh Sojoudi, Chair

Matrix sensing represents a critical, non-convex challenge within the domain of mathematical optimization, distinguished by its wide-ranging practical applications—such as medical imaging, recommender systems, and phase retrieval—as well as its significant theoretical contributions, particularly its equivalence to training a two-layer quadratic neural network. The ability to efficiently solve this problem to optimality promises substantial benefits not only for its direct applications but also provides a crucial benchmark that aids in navigating the increasingly intricate non-convex landscapes characteristic of contemporary machine learning systems. While prior research predominantly focuses on scenarios abundant in observations and characterized by a low Restricted Isometry Property (RIP) constant, thereby facilitating optimal solutions through either convex relaxation methods, including nuclear-norm minimization, or local search strategies applied to the Burer-Monteiro factorized formulation—thereby accelerating computational processes without compromising performance guarantees—the research to date remains incomplete. This is particularly true in real-world settings where acquiring a large volume of observations is often impractical, thus rendering these guarantees inapplicable.

In this dissertation, we propose innovative strategies, models, and conceptual frameworks aimed at addressing the matrix sensing problem under conditions of limited observations and noise corruptions, with the objective of provably reconstructing the ground truth matrix. Our discussion begins by exploring various methodologies of over-parametrization as a means to solve this problem, followed by an examination of alternative solutions in scenarios where over-parametrization is not used. Additionally, we delve into the impact of noise on the extraction of a global solution, offering insights into how it affects the overall process. This work serves not only as an elaborate guide to resolving matrix sensing and, by extension, low-rank optimization problems in less than ideal conditions but also endeavors to enhance our understanding of the complexities involved in non-convex optimization, thereby contributing to the broader field of mathematical optimization and machine learning.

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation stands as a beacon of my academic journey, a journey that would have been insurmountable without the collective support, wisdom, and encouragement of numerous individuals who have been part of this incredible voyage.

At the forefront of this journey is my wonderful advisor, Professor Somayeh Sojoudi. Her mentorship has been nothing short of transformational, shaping me not only as an academic but as a thinker capable of pushing the boundaries of knowledge. Professor Sojoudi's unparalleled support, coupled with her ability to create a nurturing yet challenging research environment, has been pivotal in my development. Her insightful guidance, unwavering patience, and motivating presence have illuminated my path towards achieving a clear and purposeful research direction. The essence of this dissertation is imbued with her wisdom, and it is her mentorship that has made my PhD experience not only possible but profoundly enjoyable.

Equally deserving of my deepest gratitude is Professor Javad Lavaei, whose collaboration has been a cornerstone of my PhD journey. His mentorship extended beyond conventional boundaries, introducing me to the intricate world of low-rank matrix problems and the challenging landscapes of non-convex optimization. Our extensive, thought-provoking discussions have been instrumental in refining my research, pushing me to explore and validate my hypotheses with greater depth. Professor Lavaei's brilliant insights and innovative ideas have been a wellspring of inspiration, fostering numerous projects that have enriched this dissertation. His role in my academic development is immeasurable, serving not just as a mentor but as a role model in my scholarly pursuits.

I am also thankful to my collaborators Yingjie Bi, Ying Chen, Igor Molybog, Baturalp Yalcin, Brendon Anderson, and Jingqi Li. I am profoundly grateful for their willingness to share their knowledge and for the synergy that our collaborations engendered.

My journey at Berkeley was also shaped by the support of my friends and colleagues—Fangda Gu, Jingqi Li, Yatong Bai, Brendon Anderson, Zhuohan Li, Zhiyang He, Yaodong Yu, Donghao Ying, Haixiang Zhang, Salar Fattahi, Samuel Pfrommer, Elizabeth Glista, Eli Brock, and Hyunin Lee. Their support, especially during the difficult times imposed by COVID restrictions, made my journey at Berkeley much easier. Special thanks to Haixiang Zhang and Salar Fattahi, whose insightful discussions and shared wisdom added substantial value to my research endeavors.

Lastly, the unconditonal support of my family and friends has been the cornerstone of my resolve and perseverance. I would particularly like to acknowledge Dingyi Yu, Shaolong Tian, Tianchang Shen, Haowei Zhang, Bingran Wang, and Chuhan Zhou for bringing joy and happiness into my life during my time at Berkeley. I must also express my deepest gratitude to my parents, Qianyu Zhao and Xuguang Ma, for always being my strongest supporters. My journey is as much theirs as it is mine, and I thank them with a heart full of love.

Chapter 1

Introduction

1.1 The Low-Rank Recovery Problem

In this dissertation, we mostly focus on the important problem of matrix sensing, albeit with sections that extends beyond this specific problem and applies to a wider range of low-rank recovery problems, which matrix sensing is a part of. Therefore, as a preluding argument, it is beneficial to talk about the general low-rank recovery problem to start off our discussion.

Low-rank matrix recovery has both direct and indirect applications across various fields, exploiting the fact that many real-world data structures can be represented as low-rank matrices. Direct applications often involve solving problems where the low-rank structure of the matrix is a central aspect of the data itself or the problem to be solved. Indirect applications, meanwhile, utilize low-rank matrix recovery as an intermediate step to facilitate or enhance other processes or analyses. Direct applications include:

1. **Principal Component Analysis (PCA)**: PCA is a foundational technique in data science and engineering for dimensionality reduction, enabling the simplification of data to its most informative components. Low-rank matrix recovery underpins PCA by extracting the underlying low-dimensional structure from high-dimensional datasets. This approach is particularly effective against large errors in structured data, even when traditional PCA fails due to its sensitivity to sparse, high-magnitude errors [12, 81, 80].

2. **Matrix Completion and Sensing**: These techniques are pivotal in scenarios where the goal is to infer or reconstruct a matrix from a subset of its entries. Matrix completion finds extensive applications in collaborative filtering and recommender systems, where it aids in predicting user preferences with minimal initial information. Similarly, matrix sensing is crucial in signal processing and compressed sensing, where it assists in recovering signals or images from incomplete or corrupted measurements [15, 16, 72]

3. **Computer Vision and Image Processing**: Low-rank matrix recovery is used to address problems like image compression, noise reduction, and feature extraction [89, 85, 23].

whereas indirect applications also play important roles in many scenarios:

1. **Machine Learning and Signal Processing**: In machine learning, low-rank approximation is employed to simplify models, reduce overfitting, and enhance computational efficiency. By approximating high-dimensional data with a low-rank structure, it is possible to speed up algorithms and make them more interpretable. Signal processing benefits similarly, using low-rank matrices to filter noise from signals, thereby improving the quality and reliability of the data [84, 90]. Notably, recent advances in large languages models (LLM) also take advantage of the nature of low-rank matrices to accelerate fine-tuning [30].

2. **Network Analysis**: Low-rank matrix techniques are also instrumental in analyzing complex networks, such as social networks or biological systems. They help uncover hidden patterns, predict connections, or deduce the state of a network from incomplete observations, thereby providing insights into the structure and dynamics of these systems [50].

If we focus on the applications of low-rank matrix recovery problems in machine learning and data analytics, which is where we hope to apply our new results to, it also has numerous applications, including collaborative filtering [42], phase retrieval [75, 8, 74], motion detection [23], and power system state estimation [104, 36]. To get a better grasp of the problem, we formally define it as follows: Given a measurement operator $\mathcal{A}(\cdot) : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^m$ returning a m-dimensional measurement vector $\mathcal{A}(M^*)$ from a low-rank ground truth matrix $M^* \in \mathbb{R}^{n_1 \times n_2}$ with rank r, the goal is to obtain a matrix with rank less than equal to r that conforms with the measurements, preferably the ground truth matrix M^*. This problem can be stated as the feasibility problem

$$\begin{aligned} \text{find} \quad & M \in \mathbb{R}^{n \times n} \\ \text{s.t.} \quad & \mathcal{A}(M) = \mathcal{A}(M^*) \\ & \operatorname{rank}(M) \leq r. \end{aligned} \tag{1.1}$$

While the measurement operator $\mathcal{A}$ can be nonlinear as in the case of one-bit matrix sensing [20] and phase retrieval [74], matrix sensing and matrix completion that are widely studied have linear measurement operators [15, 72]. We focus on the matrix sensing and matrix completion problems throughout this paper. Despite the linearity of $\mathcal{A}$, there are two types of problems depending on the structure of the ground truth matrix M^*. The first type, symmetric problem, consists of a low-rank positive semidefinite ground truth matrix $M^* \in \mathbb{R}^{n \times n}$, whereas the second type, asymmetric problem, consists of a ground truth matrix $M^* \in \mathbb{R}^{n_1 \times n_2}$ that is possibly sign indefinite and non-square. Since each asymmetric problem can be converted to an equivalent symmetric problem [94], we study only the symmetric problem in this paper.

The matrix sensing and completion problems have linear measurements; hence, the first constraint in problem (1.1) is linear. Therefore, the only nonconvexity of the problem arises from the nonconvex rank constraint. Earlier works on these problems focused on their convex relaxations by penalizing high-rank solutions [15, 72, 16]. They utilized the nuclear norm of a matrix as the convex surrogate of the rank function. This led to semidefinite programming (SDP) relaxations, which solve the original non-convex problems exactly with high probability based on some assumptions on the linear measurement operator and the ground truth matrix, such as the Restricted Isometry Property (RIP) and incoherence conditions. High computational time and storage requirements of the SDP algorithms incentivized the implementation of the Burer-Monteiro (BM) factorization approach [10]. This approach factorizes the symmetric matrix variable $M \in \mathbb{R}^{n\times n}$ as $M = XX^T$ for some matrix $X \in \mathbb{R}^{n\times r}$, which obviates imposing the positive semi-definiteness and rank constraints. Although the dimension of the decision variable reduces dramatically when r is small, the problem is still non-convex since its objective function is non-convex in terms of the factorized X.

1.2 Matrix Sensing Problems

In this section, we formally introduce the matrix sensing problem, because this is the centerpiece of study in this dissertation. As explained in the previous section, matrix sensing is a special case of low-rank matrix recovery, in the sense that we only consider linear measurement. When we talk about linear measurements, we are almost exclusively talking about the following linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$:

$$\mathcal{A}(M) = [\langle A_1, M\rangle, \langle A_2, M\rangle, \dots, \langle A_m, M\rangle]^T \quad \forall M \in \mathbb{R}^{n\times n}$$

where $\{A_i\}_{i=1}^m \in \mathbb{R}^{n\times n}$ are called sensing matrices. They could also be assumed to be symmetric without loss of generality, because we could simply replace A_i with $(A_i + A_i^\top)/2$ without changing the measurements. It is important to note that in all matrix sensing problems we have access to all sensing matrices. Generally speaking, there are two ways of dealing with matrix sensing problem, 1) solving via convex semi-definite programming (SDP) relaxation and 2) solving it directly with BM factorization.

Formally, the SDP formulation of the matrix sensing problem uses the nuclear norm of the variable, $\|M\|_*$, to serve as a surrogate of the rank, and replaces the rank constraint in (1.1) with an objective to minimize $\|M\|_*$. Due to the symmetricity and positive semidefiniteness of the variable, the nuclear norm is equivalent to the trace of the matrix variable M. Hence, the SDP formulation can be written as

$$\min_{M\in\mathbb{R}^{n\times n}} \quad \operatorname{tr}(M) \qquad \text{s.t. } \mathcal{A}(M)=b,\ \mathbf{M}\succeq 0, \tag{1.2}$$

where $b=\mathcal{A}(M^*)$ is given. Moreover, the matrix completion problem is a special case of the matrix sensing problem with each sensing matrix measuring only one entry of M^*. We can represent the measurement operator $\mathcal{A}$ as $\mathcal{A}_\Omega : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n}$ for this special case, which is defined as follows:

$$\mathcal{A}_\Omega(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij} & \text{if } (i,j)\in\Omega \\ 0 & \text{otherwise,} \end{cases}$$

where Ω is the set of indices of observed entries. Although being convex, the main drawback of (1.2) is that its global solution might not be M^*, the ground truth solution we hope to recover. Therefore we could always reach the global solution of (1.2), but it might be a matrix totally unrelated to the ground truth, and might even possess a higher rank than M^*.

Besides the SDP formulation, the BM factorization formulation of matrix sensing is attracting increasing attention from the community due to its simplicity in form and also its reduced computational complexity when compared to the SDP formulation (1.2). This BM approach will also be the main subject of study in this piece due to its non-convexity. The optimization landscape of non-convex problems is notoriously complex to analyze in general due to the existence of an arbitrary number of spurious solutions (a spurious solution is a second-order critical point that is not a global minimum). As a result, if a numerical algorithm is not initialized close enough to a desirable solution, it may converge to one of those problematic spurious solutions. It may be acceptable (depending on the application) if the algorithm finds a critical point different from but close to the true solution, while converging to a point faraway implies the failure of the algorithm. Therefore, as we further the study of matrix sensing in its factorized form, we also get better ideas of what non-convex landscapes look like in general and how to better solve them. Formally stated, this problem is defined as:

$$\min_{X\in\mathbb{R}^{n\times r_{\text{search}}}} h(X) := \frac{1}{2}\|\mathcal{A}(XX^T)-b\|^2 \tag{1.3}$$

where $b=\mathcal{A}(M^*)$ is a measurement of our desired measurement. Here, we define the second dimension of X to be the search rank r_{search}, which might differ from M^*'s true rank r. This is because in many applications we do not have prior access to r, therefore requiring us to overestimate it with $r_{\text{search}} \geq r$, a practice we usually refer to as ”over-parametrization”. More recently, it is also discovered that such over-parametrization can lead to better optimization landscapes [97, 76, 51]. Note again for this dissertation we maintain the assumption that we hope to recover *symmetirc* matrices. However, all the techniques presented in this dissertation will also apply to non-symmetric matrices since there exist tricks to do so under this factorized formulation. We leave the details of this conversion to Section 2.2. For the sake of notational convenience, we further define the function $f(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}$:

$$f(M) := \frac{1}{2}\|\mathcal{A}(M) - b\|^2 \quad \forall M \in \mathbb{R}^{n\times n} \implies f(XX^\top) = h(X) \tag{1.4}$$

The major drawback of (1.3) is that it may have an arbitrary number of spurious solutions, which cause ubiquitous local search algorithms to potentially end up with unwanted solution. Therefore, there has been an extensive investigation of the non-convex optimization landscape of (1.3), and the centerpiece notion is the restricted isometry property (RIP), defined below.

Definition 1 (RIP [15]) *Given a natural number p, the linear map $\mathcal{A} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ is said to satisfy δ_p-RIP if there is a constant $\delta_p \in [0, 1)$ such that*

$$(1 - \delta_p)\|M\|_F^2 \leq \|\mathcal{A}(M)\|^2 \leq (1 + \delta_p)\|M\|_F^2$$

holds for matrices $M \in \mathbb{R}^{n\times n}$ satisfying $\operatorname{rank}(M) \leq p$.

Intuitively speaking, a smaller RIP constant means that the problem is easier to solve. For instance, if $\delta_{2r} = 0$, then $\mathcal{A}(\cdot)$ becomes the identity operator with $b = \operatorname{vec}(M^*)$, which makes the problem trivial to solve for M^*. Traditionally, the RIP constant is tightly tied with the number of observations m under the assumption that the sensing matrices are sampled randomly, most commonly from Gaussian distribution. In the classic work [14], the authors showed that $\mathcal{O}(1/\delta^2)$ number of random Gaussian measurements are required to ensure δ-RIP_{2r}. Therefore to achieve a good (low) RIP constant, large number of measurements are required, which might not be accessible. Additionally, for certain parts of this dissertation, we might adopt an equivalent characterization to RIP, which are the Restricted Strong Smoothness (RSS) and Restricted Strong Convexity (RSC) constants. This alternative definition better captures the asymmetry in the upper and lower bounds, because using one variable δ_p might not capture that asymmetry. Here we formally introduce it

Definition 2 (RSS and RSC) *The linear operator $\mathcal{A} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ satisfies the (L_s, p)-RSS property and the (α_s, p)-RSC property if*

$$\begin{aligned} f(M) - f(N) &\leq \langle M - N, \nabla f(N)\rangle + \frac{L_s}{2}\|M - N\|_F^2 \\ f(M) - f(N) &\geq \langle M - N, \nabla f(N)\rangle + \frac{\alpha_s}{2}\|M - N\|_F^2 \end{aligned}$$

are satisfied, respectively for all $M, N \in \mathbb{R}^n$ with $\operatorname{rank}(M), \operatorname{rank}(N) \leq p$. Note that RSS and RSC provide a more expressible way to represent the RIP property, with $\delta_p = (L_s - \alpha_s)/(L_s + \alpha_s)$.

In the next section, we provide a review of how RIP plays a central role in determining the optimization landscape of the non-convex problem (1.3), and explain why this problem still needs a further investigation even with the abundance of literature dedicated to this topic. To further streamline our presentation, we divide our discussion into two parts: 1) Ideal scenarios (noiseless and RIP being smaller than 1/2) and 2) Non-ideal cases (with noise, or RIP being greater than 1/2). By discussing the prior works in this fashion, we hope to convey to the readers the necessities of building guarantees and techniques to deal with less ideal cases, especially because these less ideal cases align much better with real-life problems and could offer us better insights into non-convex optimization in general.

1.3 Related Works

The matrix sensing and matrix completion problems have been extensively studied, with a significant focus on understanding the optimization landscape of the non-convex Burer-Monteiro (BM) formulation and the convex semidefinite programming (SDP) relaxation. In this section, we review the related works, organized into three main categories: (1) the ideal regime where the restricted isometry property (RIP) constant is smaller than 1/2 and no noise exists, (2) the regime where the RIP constant is larger than 1/2, and (3) the regime where our measurements are corrupted by noise.

1.3.1 Ideal Cases

The attention to the RIP constant was first popularized by the study of using a convex semidefinite programming (SDP) relaxation to solve the matrix sensing problem [72, 16]. It was proven that as along as $\delta_{5r} \leq 1/10$, the SDP relaxation was tight and M^* could be recovered exactly.[11] later improved this bound to 1/2 for the convex SDP relaxation formulation.

Subsequently, [5, 24] analyzed the factorized problem (1.3) and concluded that as long as $\delta_{2r} \leq 1/5$, all second-order critical points (SOPs) of (1.3) are ground truth solutions. [105, 46] also proved that $\delta_{4r} \leq 1/5$ is sufficient for the global recovery of M^* under an arbitrary objective function (instead of the least-squares one in (1.3)). Later, by using a ”certification of in-existence” technique, [101] established that $\delta_{2r} = 1/2$ was a sharp bound when we constrain our search rank to be r, meaning that as long as $\delta_{2r} < 1/2$, all problem instances of (1.3) are free of spurious solutions, and once $\delta_{2r} \geq 1/2$, it is possible to establish counter-examples with SOPs not corresponding to ground truth solutions [94]. This aforementioned approach is important because it quantifies how restrictive RIP needs to be in order to ensure a benign landscape.

Furthermore, when the RIP constant is small enough, various desirable properties hold, including fast convergence [48, 82] and spectral contraction [76, 34].

1.3.2 RIP Constant Larger than $1/2$

When the RIP constant of the problem is larger than or equal to $1/2$, the optimization landscape of the BM problem becomes highly non-convex, and counterexamples can be found with SOPs that are not global solutions [94]. Few works have attempted to provide limited mathematical guarantees in this regime.

Benign Landscape Near M^*:[101] proved that when $\delta_{2r} \geq 1/2$ for $r = 1$, the absence of spurious solutions can be ensured in a local region close to M^*, depending on the RIP constant and the size of M^**. [93] expanded this analysis to the general r case, demonstrating the ubiquity of this phenomenon.

Over-parametrization with $r_{\textbf{search}} > r$: This line of work investigates the case where the search rank r_{search} is greater than the true rank r, leading to increased algorithmic complexity. [97] proved that if $r_{\text{search}} > r[(1+\delta_n)/(1-\delta_n)-1]^2/4$, with $r^* \leq r < n$, then every SOP $\hat{X}$ satisfies $\hat{X}\hat{X}^\top = M^*$. [51] derived a similar result for the ℓ_1 loss under an RIP-type condition. Despite the superiority of guarantees since the bound can go over $1/2$, the power of the stated over-parametrization is limited. The reason is that r_{search} cannot be greater than n and therefore it is impossible to satisfy the condition in practical cases where RIP constant is large. This calls for a new framework that accommodates an arbitrarily large degree of parametrization.

1.3.3 Noisy Environments

By noisy environments, we refer to the case where our measurement b is influenced by noise and become $\tilde{b}$ with

$$\tilde{b} = b + w, \quad w \sim \mathcal{D}, \mathbb{E}[\|w\|^2] < \infty$$

with w being sampled from some finite-variance family $\mathcal{D}$ that admits valid concentration inequalities. For the noisy problem, the relation $ZZ^\top = M^*$ is unlikely to be satisfied, where Z denotes a global minimizer of problem (1.3). However, in this situation, $ZZ^\top$ should be close to the ground truth M^* if the noise w is small. As a generalization of the above-mentioned results for the noiseless problem, it is natural to study whether all local minimizers, including the global minimizers, are close to the ground truth M^* under the RIP assumption. One such result is presented in [5] and given below.

Theorem. *Suppose that $w \sim \mathcal{N}(0, \sigma_w^2 I_m)$ and $\mathcal{A}(\cdot)$ has the δ-RIP$_{4r}$ property with $\delta < 1/10$. Then, with probability at least $1-10/n^2$, any local minimizer $\hat{X}$ of problem* (1.3) *satisfies the inequality*

$$\|\hat{X}\hat{X}^T - M^*\|_F \leq 20\sqrt{\frac{\log(n)}{m}}\sigma_w.$$

Theorem 31 in [24] further improves the above result by replacing the δ-RIP$_{4r}$ property with the δ-RIP$_{2r}$ property. [47] studies a similar noisy low-rank matrix recovery problem with l_1 norm.

Furthermore, [102] proves that all local minima are close to the ground truth when $\delta \leq 1/35$ for a general objective, which is an extremely strong assumption on δ. Furthermore, [102] requires the RIP condition to be satisfied for the noisy problem rather than its noiseless counterpart, which is impossible to verify beforehand due to the unknown noise.

1.4 Summary of Contributions

This dissertation centers on demystifying the solving of this non-convex problem of matrix sensing under more realistic settings. By realistic settings, we mostly focus on two scenarios, 1) the under-sampled regime, where the RIP constant is high, and 2) when the observation is complicated by random noise.

1.4.1 Under-Sampled Regime

This regime represents a very realistic setting where we only have access to a small number of observations, which under a random Gaussian assumption, leads to high RIP constants. This regime also encompasses cases where the measurement matrices are deterministic, and are known to have high RIP constant, like those mentioned in [88]. Given the literature review given above, we have the basic understanding that when $\delta_{2r} < 1/2$, both the SDP approach (1.2) and the BM approach (1.3) have nice gaurantees. The SDP approach can directly recover M^*, while the BM landscape is free of spurious solutions, paving way for saddle-escaping algorithms to reach the global solution in polynomial time [33]. Therefore, this dissertation specifically investigates improved guarantees and new strategies to tackle the matrix sensing problem when $\delta_{2r} \geq 1/2$, in the hope to get better solve this problem to optimality.

Chapter 3, Improved SDP Guarantee:

Papers like [100, 94] showcased that when using the BM factorization with $r_{\text{search}} = r$, δ_{2r} was a sharp threshold, meaning that instances of matrix sensing can be found with spurious solutions as soon as $\delta_{2r} = 1/2$, giving practitioners a clear-cut line. However, when it comes to the SDP approach, the best bound was only proven to be sufficient, and not necessary. Given the more computationally expensive nature of this approach, we hope to better its theoretical guarantees to see if it has improved guarantees compared to the BM approach. Indeed, in this dissertation, we prove that there exists a lower bound δ_{lb} on the RIP constant δ to guarantee convergence to the ground truth solution by using a proof technique called the in-existence of incorrect solution [101]. We aim to find a linear measurement operator $\mathcal{A}$ with the smallest RIP constant such that the SDP formulation converges to a wrong solution. We found that contrary to the BM method which exhibits the same performance independent of the rank of the unknown solution, the success of the SDP method is correlated to the rank of the solution and improves as the rank increases. This bound δ_{lb} can go well over the previously known tightest 1/2, and could even approach 1 when $n \approx 2r$.

This chapter is mostly based on part of this AAAI'23 oral paper [88].

Chapter 4, Lifted Framework via Tensors:

In the previous chapter, we have proved that by using the SDP approach, which is inherently an over-parametrized model, we could achieve better guarantees than the exact-parametrized BM model. Over-parametrized BM with $r_{\text{search}} \geq r$ like those mentioned in Section 1.3.2 offer a nice starting point, but still lacks applicability in certain cases. Therefore, in this chapter we propose an innovative over-parametrization technique that lifts our search space from matrices to tensors. This can also be seen as a way to apply Burer-Monteiro factorization on various levels of SDP problems induced by dual of Sum-of-Squares (SOS) optimization. This contrasts with the existing over-parametrization technique where the search rank is limited by the dimension of the matrix and it does not allow a rich over-parametrization of an arbitrary degree. We show that although the spurious solutions of the problem remain stationary points through the hierarchy, they will be transformed into strict saddle points (under some technical conditions) and can be escaped via local search methods. We also derive a bound on how much over-parametrization is required to enable the elimination of spurious solutions. Furthermore, we show that with sufficiently small initialization scale, gradient descent applied to this lifted problem results in approximate rank-1 tensors and critical points with escape directions. Our findings underscore the significance of the tensor parametrization of matrix sensing, in combination with first-order methods, in achieving global optimality in such problems.

This chapter is based on this ICML'23 oral paper [56] and this NeurIPS'23 paper [55].

Chapter 5, Higher-Order Loss Function:

In the previous two chapters, we talked about how an increased parametrization can lead to better guarantees in reaching the ground truth matrix. However, due to many real-world constraint, it is possible that such over-parametrization might not be possible in many cases. Thus, in this chapter, we prove that under certain conditions, critical points sufficiently distant from the ground truth matrix exhibit favorable geometry by being strict saddle points rather than troublesome local minima. Moreover, we introduce the notion of higher-order losses for the matrix sensing problem and show that the incorporation of such losses into the objective function amplifies the negative curvature around those distant critical points. This implies that increasing the complexity of the objective function via high-order losses accelerates the escape from such critical points and acts as a desirable alternative to increasing the complexity of the optimization problem via over-parametrization. By elucidating key characteristics of the non-convex optimization landscape, this work makes progress towards a comprehensive framework for tackling broader machine learning objectives plagued by non-convexity.

This work is based on this AISTATS'24 oral paper [54].

1.4.2 Noisy Regime

This regime represents another realistic consideration that our observations, irrespective of its quantity m, could be affected by random noise. This scenario is relevant in applications such as state estimation, which is crucial for the functioning of power grids and can be conceptualized through matrix sensing [36]. In this context, each data point is derived from a physical device, and the incorporated noise accounts not only for errors inherent to the sensors but also for discrepancies between the actual system's behavior and its theoretical model, alterations due to cyber-attacks, mechanical failures, and more. From a mathematical standpoint, when we apply the BM approach, we encounter a problem that is slightly altered by noise, represented as:

$$\min_{X \in \mathbb{R}^{n \times r_{\text{search}}}} h_w(X) := \frac{1}{2}\|\mathcal{A}(XX^T) - \tilde{b}\|^2, \quad \tilde{b} = b - w, \quad w \sim \mathcal{D}$$

where $\mathcal{D}$ denotes any distribution with finite variance. The introduction of noise complicates matters by potentially shifting the global solution of Equation (6.1) away from the ground truth matrix M^*. This indicates that reaching the global solution does not guarantee proximity to the desired matrix. Under this critical premise, the dissertation demonstrates that the same RIP constant, which prevents the emergence of spurious solutions in the absence of noise, ensures that every second-order point (SOP) lies within a narrow margin of M^*, the sought-after true solution.

Chapter 6, Noisy Matrix Sensing:

In this chapter, we propose a global guarantee on the maximum distance between an arbitrary local minimizer and the ground truth under the assumption that the RIP constant is smaller than 1/2. We show that this distance shrinks to zero as the intensity of the noise reduces. Our new guarantee is sharp in terms of the RIP constant and is much stronger than the existing results. We then present a local guarantee for problems with an arbitrary RIP constant, which states that any local minimizer is either considerably close to the ground truth or far away from it. Next, we prove the strict saddle property, which guarantees the global convergence of the perturbed gradient descent method in polynomial time. The developed results demonstrate how the noise intensity and the RIP constant of the problem affect the landscape of the problem. Moreover, we further extend these results to the over-parametrized regime, where $r_{\text{search}} \geq r$.

This chapter is based on this AAAI'22 paper [52] and its journal extension published at the INFORMS Journal on Optimization [53].

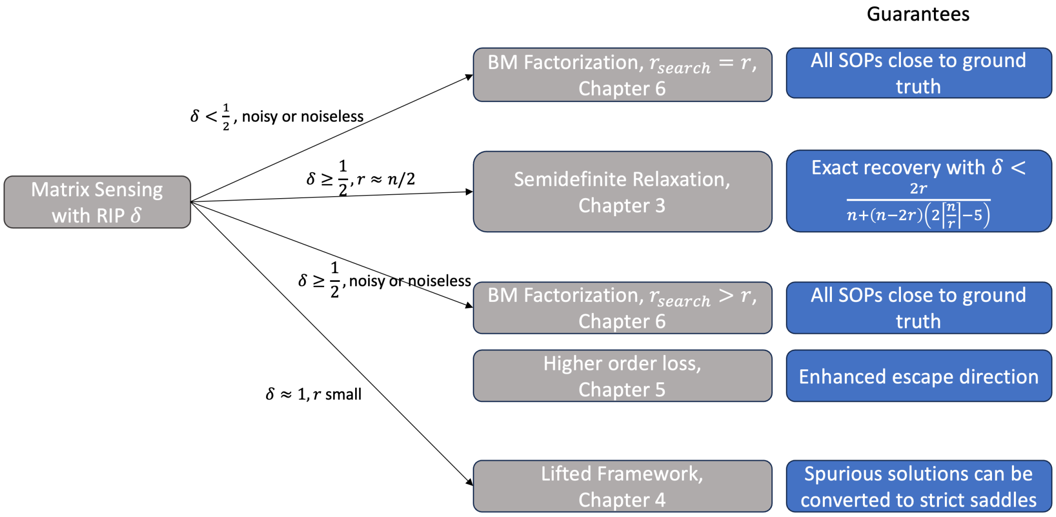
Chapter 7, Noisy General Low-Rank Optimization:

This chapter serves as a further extension of the previous chapter, in which we discuss the effects of noise on a more general low-rank recovery problem introduced as (1.1). In this chapter, we introduce new guarantees on solving the general problem (1.1) with BM factorization with a far less restrictive RIP constant. We prove that as long as the RIP constant of the noiseless objective is less than $1/3$, any spurious local solution of the noisy optimization problem must be close to the ground truth solution. By working through the strict saddle property, we also show that an approximate solution can be found in polynomial time. We characterize the geometry of the spurious local minima of the problem in a local region around the ground truth in the case when the RIP constant is greater than $1/3$.

This chapter is based on this AISTATS'23 oral paper [57].

1.4.3 Overall Framework

The series of results introduced above gave us new approaches to solving matrix sensing in more realistic settings and outlined their respective guarantees. Combined with the previous works in this field, we now have a rather complete picture of how to solve matrix sensing problems in general, outlined in this following figure:


Guarantees
Matrix Sensing with RIP δ
$\delta < \frac{1}{2}$, noisy or noiseless
$\delta \geq \frac{1}{2}, r \approx n/2$
$\delta \geq \frac{1}{2}$, noisy or noiseless
$\delta \approx 1, r$ small
BM Factorization, $r_{search} = r$, Chapter 6
Semidefinite Relaxation, Chapter 3
BM Factorization, $r_{search} > r$, Chapter 6
Higher order loss, Chapter 5
Lifted Framework, Chapter 4
All SOPs close to ground truth
Exact recovery with $\delta < \frac{2r}{n+(n-2r)\left(2\left\lceil\frac{n}{r}\right\rceil-5\right)}$
All SOPs close to ground truth
Enhanced escape direction
Spurious solutions can be converted to strict saddles

Chapter 2

Mathematical Prior

2.1 Notation

- **Matrices and Vectors:**
  - I_n: The identity matrix of size $n \times n$.
  - $M \succeq 0$: Denotes that M is a symmetric and positive semidefinite matrix.
  - $\sigma_i(M)$: The i-th largest singular value of matrix M.
  - $\lambda_i(M)$: The i-th largest eigenvalue of matrix M.
  - $\sigma_{\min}(M)/\sigma_{\max}(M)$: The least/largest non-zero singular value of M.
  - $\lambda_{\min}(M)\lambda_{\max}(M)$: The least/largest non-zero eigenvalue of M.
  - $\|v\|$: The Euclidean norm of vector v.
  - $\|M\|_F$ and $\|M\|_p$: The Frobenius norm and induced l_p norm of matrix M, respectively, for $p \geq 2$. If unspecified, we default $\|M\| = \|M\|_2$
  - $\langle A, B \rangle = \mathrm{tr}(A^\top B)$: The inner product of matrices A and B of the same size.
  - $A \oslash B$: The Kronecker product of matrices A and B.
  - $\circ l$ stands for the shorthand of repeated cartesian product $\times \cdots \times$ for l times.
  - $\otimes$ denotes tensor outer product.
- **Vectorization and Matrix Operations:**
  - $\mathrm{vec}(M)$: The vectorization of matrix M, stacking its columns into a vector.
  - $\mathrm{mat}(v)$: Converts a vector $v \in \mathbb{R}^{n^2}$ to a square matrix.
  - $\mathrm{mat}_S(v) = (M + M^\top)/2$: Converts v to a symmetric matrix, where M satisfies $v = \mathrm{vec}(M)$.

- **Statistical Distributions:**
  - $\mathcal{N}(\mu, \Sigma)$: The multivariate Gaussian distribution with mean μ and covariance matrix Σ.
- **Differential Calculus:**
  - $\nabla f(\cdot)$ and $\nabla^2 f(\cdot)$: The gradient and Hessian of a function $f : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$, respectively.
  - For a function with matrix as input $f(\cdot) : \mathbb{R}^{n_1\times n_2} \mapsto \mathbb{R}$, its Hessian is a four-dimensional tensor, with the notation $[\nabla^2 f(\mathbf{X})]_{i,j,k,l} = \frac{\partial^2 f(\mathbf{X})}{\partial \mathbf{X}_{i,j} \partial \mathbf{X}_{k,l}}$. This Hessian can be regarded as a quadratic form whose action on any two matrices $K, L \in \mathbb{R}^{n_1\times n_2}$ is given by

$$[\nabla^2 f(\mathbf{X})](K, L) = \sum_{i,j,k,l=1} \frac{\partial^2 f}{\partial \mathbf{X}_{ij} \partial \mathbf{X}_{kl}} (\mathbf{X}) K_{ij} L_{kl}.$$

- **Miscellaneous:**
  - $[n]$: The integer set $\{1, \dots, n\}$.
  - $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$: The ceiling and floor operators, respectively.
  - $|\mathcal{S}|$: The cardinality of set $\mathcal{S}$.
  - $\mathbb{S}^n_+$: The set of $n \times n$ positive semidefinite matrices.
  - $\asymp$ denotes "asymptotic to", meaning that the two terms on both sides of this symbol have the same order of magnitude.

We also characterize the distance of an arbitrary factorized point $X \in \mathbb{R}^{n\times r}$ to a rank-r positive semidefinite matrix M with the function $\operatorname{dist}(X, M)$, defined as:

$$\operatorname{dist}(X, M) = \min_{Z\in\mathcal{Z}} \|X - Z\|_F,$$
$$\mathcal{Z} = \{Z \in \mathbb{R}^{n\times r} \mid M = ZZ^\top\}.$$

Given a matrix $\hat{X} \in \mathbb{R}^{n\times r}$, define $\hat{\mathbf{X}} \in \mathbb{R}^{n^2\times nr}$ to be the matrix satisfying

$$\hat{\mathbf{X}} \operatorname{vec}(U) = \operatorname{vec}(\hat{X} U^\top + U \hat{X}^\top), \quad \forall U \in \mathbb{R}^{n\times r}.$$

Define $\mathcal{P}_r(M)$ of an arbitrary matrix M to be the projection of M on a low-rank manifold of rank at most r:

$$\mathcal{P}_r(M) = \operatorname*{arg\,min}_{M_r\in\mathcal{M}} \|M_r - M\|_F,$$
$$\mathcal{M} := \{M \in \mathbb{S}^{n\times n} | \operatorname{rank}(M) \le r, M \succeq 0\}$$

Lastly, in some cases, the bolded $\mathbf{A} \in \mathbb{R}^{m\times n^2}$ is defined such that $\mathbf{A} \operatorname{vec}(M) = \mathcal{A}(M)$.

2.2 Useful Tools

General Identities

In this section we provide some basic algebraic facts for matrices that will aid the understanding of our results. For a more complete review, please refer to this universally acclaimed book [67].

- If $A, B \succeq 0$ are both PSD matrices, then we have
$$\langle A, B \rangle \geq \lambda_{\min}(A) \operatorname{tr}(B)$$
- For an arbitrary matrix $A \in \mathbb{R}^{n_1 \times n_2}$, and two vectors $x \in \mathbb{R}^{n_1}, y \in \mathbb{R}^{n_2}$, we have
$$\langle A, xy^\top \rangle = x^\top A y$$
- For any arbitrary matrix A, $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$, called the *spectral norm.*
- For any arbitrary matrix $A \in \mathbb{R}^{n_1 \times n_2}$,
$$\|A\|_F^2 = \sum_{i,j}^{n_1,n_2} A_{i,j}^2 = \|\operatorname{vec}(A)\|_2^2$$
- For any orthonormal basis $U \in \mathbb{R}^{n \times k}$ and any vector $x \in \mathbb{R}^k$, $\|Ux\|_2 = \|x\|_2$.
- For any orthonormal basis $U \in \mathbb{R}^{n \times k}$ we have $\|U\|_F \leq \sqrt{k}$.
- For symmetric matrix A, we have $\sigma_{\min}(A) = \min_{\|z\|_2=1} z^\top A z$ and $\|A\|_2 \geq z^\top A z$.
- For any matrix A, we have $\|A\|_2 \leq \|A\|_F$
- For any square and invertible matrix $R \in \mathbb{R}^{n \times n}$, we have $\|R^{-1}\|_2 = \sigma_{\min}(R)^{-1}$.
- For any matrix $A \in \mathbb{R}^{n_1 \times n_2}$ and vector $x \in \mathbb{R}^{n_2}$, we have that $\|A\|_2 \geq \|Ax\|_2$.
- For any matrix $A \in \mathbb{R}^{n_1 \times n_2}$,
$$\|AA^\top\|_2 = \|A^\top A\|_2$$
- For any vector $x \in \mathbb{R}^n$,
$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$$

Matrix Sensing Facts

There are some other useful identities specifically tied to the BM formulation of matrix sensing (1.3). To be more specific, we could put bounds on the r^{th} singular value of critical points of (1.3). This property will prove itself to be very useful, and appearing multiple times in this dissertation.

Lemma 1 ([57]) *For any SOP $\hat{X}$ of* (1.3) *satisfying RSS with constant L_s for arbitrary rank, define G as $G := -\lambda_{min}(\nabla f(\hat{X}\hat{X}^\top))$, and L_s be the RSS constant. Then it holds that*

$$G \le \lambda_r(\hat{X}\hat{X}^\top)L_s$$

where r is the search rank of (1.3).

Lemma 2 *Given an FOP $\hat{X}$ of* (1.3) *satisfying RSS and RSC with constants L_s, α_s for arbitrary rank, it holds that*

$$\lambda_r(\hat{X}\hat{X}^\top) < \frac{L_s}{\sqrt{r}\alpha_s}\|M^*\|_F \tag{2.1}$$

Proof 1 (Proof of Lemma 2) *Proof of Lemma 6 of [94] states that given matrix $X \in \mathbb{R}^{n\times r}$ such that it is FOP of* (1.3), *one can write*

$$0 = \langle \nabla f(XX^\top), XX^\top \rangle \ge \alpha_s \|XX^\top\|_F^2 - L_s\|M^*\|_F\|XX^\top\|_F$$

Therefore this means that

$$L_s\|M^*\|_F \ge \alpha_s\|XX^\top\|_F$$

Then realizing

$$\|\hat{X}\hat{X}^\top\|_F^2 \ge r\lambda_r(\hat{X}\hat{X}^\top)^2 \implies \lambda_r(\hat{X}\hat{X}^\top) < \frac{L_s}{\sqrt{r}\alpha_s}\|M^*\|_F$$

as $\hat{X}\hat{X}^\top$ can have at most r eigenvalues due to its factorized form.

Although this dissertation focuses on the symmetric case, meaning that the matrix of interest M is assumed to be symmetric and positive semidefinite, our analysis techniques are all valid for the case where $M \in \mathbb{R}^{n\times m}$ for arbitrary numbers m and n. To explain this generalization as per [78], we first need to deal with the redundancy of global optima induced by the asymmetry. This can be achieved by solving the following optimization problem with a regularization term instead of the original one:

$$\min_{U\in\mathbb{R}^{n\times r},\ V\in\mathbb{R}^{m\times r}} f(UV^\top, w) + \frac{\phi}{4}\|U^\top U - V^\top V\|_F^2. \tag{2.2}$$

where ϕ is an arbitrary penalization constant. As per [7], solving (2.2) is equivalent to:

$$\min_{X\in\mathbb{R}^{(n+m)\times r}} f_a(XX^\top, w) \tag{2.3}$$

where $X = \begin{bmatrix} U^\top & V^\top \end{bmatrix}^\top \in \mathbb{R}^{(n+m)\times r}$ and the function $f_a(\cdot, w) : \mathbb{R}^{(n+m)\times(n+m)} \mapsto \mathbb{R}$ satisfies:

$$f_a(\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}, w) = \frac{f(P_{11}, w) + f(P_{22}, w)}{2} + \\ \frac{\phi}{4}(\|P_{11}\|_F^2 + \|P_{22}\|_F^2 - \|P_{12}\|_F^2 - \|P_{21}\|_F^2)$$

where $P_{11} \in \mathbb{R}^{n\times n}$, $P_{12} \in \mathbb{R}^{n\times m}$, $P_{21} \in \mathbb{R}^{m\times n}$, $P_{22} \in \mathbb{R}^{m\times m}$ are just partitioned blocks of $XX^\top$ of appropriate dimensions corresponding to $UU^\top, UV^\top, VU^\top, VV^\top$, respectively. The equivalence between the asymmetric problem and its symmetric counterpart (2.3) implies that the results of this paper obtained for this dissertation can be restated for the original asymmetric problem. Note that in this section we have assumed that the objective could be noisy, and if there are no noise, one can simply set $w = 0$ and it does not change this argument at all.

2.3 Optimization Basics

Since this dissertation mostly focuses on the optimization landscape of matrix problems, we first hope to review some basic concepts about optimization theory. In general, optimization problems for a general function f can be written in this form

$$\min_{x\in E} f(x), \quad f(\cdot) : E \mapsto \mathbb{R}$$

where E is just some arbitrary Euclidean space. If this problem admits an optimal solution, we (in this review) denote it as

$$x_{\text{opt}} = \arg\min_{x\in E} f(x)$$

When the objective function is convex, then an optimal solution always exists, and furthermore if this objective function is strictly convex, it will admit a unique solution. We further note that the optimal solution may be non-unique when the objective function is non-convex. Recall that a function f is convex if and only if

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y), \quad \forall x, y \in \text{dom}(f),\ \lambda \in (0,1)$$

It is called strictly convex if the above inequality is made strict except for when $x = y$. The above definition is called the zeroth-order definition of convex functions. Equivalent first and second order definitions also exist for differentiable and twice-differentiable functions, which can be found in this book [9].

Gradient Descent for Convex Functions

Assuming that our objective function f is differentiable, the most basic and universal method to solve any optimization problem is through the use of (vanilla) gradient descent (GD) algorithm, with the update rule

$$x^{t+1} = x^t - \eta \nabla f(x^t), \quad t = 0, 1, ...$$

where x^t denotes the t^{th} iteration of this algorithm, and we assume some initial point x^0 is fed into this algorithm. Here, $\eta > 0$ is the step-size of the algorithm, or sometimes referred to as the ”learning rate” in many deep learning contexts. As an NP-hard problem, non-convex optimization in general does not have any guarantee for any algorithm to always reach to its global solution provably. However, when constrained only to convex optimization, it can be shown that the vanilla GD algorithm can converge to the solution x_{opt} at a linear rate. This *linear convergence* is said to happen if the distance of any iterate x^t to x_{opt}, $\|x^t - x_{\text{opt}}\|_2$, converges to 0 as a geometric series with the progression of t. Before giving a formal result, we introduce two new prior conditions that the function has to satisfy in order to achieve linear convergence:

Definition 3 (Strong Convexity) *A twice continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be α-strongly convex in a set $\mathcal{B}$ if*

$$\nabla^2 f(x) \succeq \alpha I_n \quad \forall x \in \mathcal{B}$$

Definition 4 (Smoothness) *A twice continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be β-smooth in a set $\mathcal{B}$ if*

$$\|\nabla^2 f(x)\| \leq \beta \quad \forall x \in \mathcal{B}$$

Now we can introduce our convergence result:

Lemma 3 *(Linear Convergence with GD) Suppose that f is α-strongly convex and β-smooth in a local set $\mathcal{B}_\zeta(x_{\text{opt}})$ around the optimum x_{opt} with*

$$\mathcal{B}_\zeta(x_{\text{opt}}) := \|x - x_{\text{opt}}\|_2 \leq \zeta, \quad x^0 \in \mathcal{B}_\zeta(x_{\text{opt}})$$

If we choose the step-size $\eta = 1/\beta$, then

$$\|x^t - x_{\text{opt}}\|_2 \leq (1 - \frac{\alpha}{\beta})^t \|x^0 - x_{opt}\|_2, \quad \forall t = 0, 1, ...$$

We omit the proof here for simplicity as it can be found in many standard optimization textbooks. A nice proof can also be found in Section 2.1 of [18]. The one thing to note about this lemma is that we did not assume f to be convex in general. We only assumed that f was strongly convex and smooth in a region around any second-order point. Therefore, this convergence analysis also applies to convergence to any local solution in non-convex objectives, provided that the landscape is benign around that local solution (strongly convex and smooth), and that the algorithm was initialized in that region.

Critical Points and Optimality Condition

Besides the important GD algorithm that is widely used and analyzed in this dissertation, another critical concept in optimization landscape is the existence of critical points. An iterative algorithm like GD often converges to one of its fixed points, no matter the objective being convex or not, and we call the associated fixed points *critical points* or first-order points (FOP) of the objective function, defined as

Definition 5 *A first-order point (stationary point, critical point) $\hat{x}$ of f is any point that satisfies*

$$\nabla f(\hat{x}) = 0$$

For readers who have heard of the "KKT" condition, the above condition is actually a subset of the KKT conditions. KKT conditions are actually necessary conditions for $\hat{x}$ to be FOPs in constrained optimization. Since in this dissertation we mostly deal with unconstrained optimization, our KKT condition is reduced to only requiring the gradient to be zero, which is also sufficient in this case.

Interestingly, for any first-order point, it can only be one of three things: 1) a local minimum, 2) a local maximum, and 3) a saddle point. If f is twice continuously differentiable (which it is under the scope of this dissertation), then any first order point $\hat{x}$ can be completed characterized by its Hessian matrix:

1. If $\nabla^2 f(\hat{x}) \succ 0$, then $\hat{x}$ is a local minimum;
2. If $\nabla^2 f(\hat{x}) \prec 0$, then $\hat{x}$ is a local maximum;
3. If $\lambda_{\min}(\nabla^2 f(\hat{x})) = 0$, then $\hat{x}$ is either a local minimum or a degenerate saddle point (saddle point with no escape direction).
4. If $\lambda_{\min}(\nabla^2 f(\hat{x})) < 0$, then $\hat{x}$ is a strict saddle point, possessing a valid escape direction.

Based on this characterization, we can further define second-order points (SOP):

Definition 6 *A second-order point $\hat{x}$ of f is any point that satisfies*

$$\nabla f(\hat{x}) = 0, \quad \nabla^2 f(\hat{x}) \succeq 0$$

We also often call second-order points as local minimums. For strongly convex functions/strictly convex functions, only one SOP exists and that is the global solution. For general convex functions, SOPs exists in a contiguous region, and from a geometric perspective, you can think of it as a region of "flatness". For non-convex objectives, infinite number of SOPs may exist, and we classify them as global solutions and spurious solutions. A **spurious solution** is a SOP that does not attain the globally minimum value. In this dissertation spurious solution is an important concept to grasp.

When restricted to the matrix sensing problem at hand, we can further drive the necessary and sufficient conditions for matrices to be FOP and SOPs of (1.3).

Lemma 4 *Given a general matrix function $f(\cdot) : \mathcal{S}_+^n \mapsto \mathbb{R}$ that takes in a symmetric matrix variable M, if we only optimize over low-rank Ms by utilizing BM factorization to explicitly factorize $M = XX^\top$, where $X \in \mathbb{R}^{n\times r_{search}}$, then the matrix $\hat{X} \in \mathbb{R}^{n\times r}$ is a second-order point (SOP) of problem*

$$\min_{X\in\mathbb{R}^{n\times r_{search}}} f(XX^\top) \tag{2.4}$$

if and only if

$$\nabla_M f(\hat{X}\hat{X}^\top)\hat{X} = 0 \tag{2.5}$$

and

$$2\langle \nabla_M f(\hat{X}\hat{X}^\top), UU^\top\rangle + [\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) \geq 0 \tag{2.6}$$

for all $U \in \mathbb{R}^{n\times r_{search}}$. Furthermore $\hat{X} \in \mathbb{R}^{n\times r_{search}}$ is a first-order point (FOP) of (2.4) *if and only if it satisfies* (2.5)*. When constrained to the case when f takes the form of* (1.4)*, therefore when* (2.5) *is equivalent to the matrix sensing problem* (1.3)*, we can further show that:*

$$\nabla f(\hat{X}\hat{X}^\top) = \sum_{i=1}^m \langle A_i, \hat{X}\hat{X}^\top - M^*\rangle A_i, \tag{2.7a}$$

$$\nabla^2 f(\hat{X}\hat{X}^\top)(\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) = \sum_{i=1}^m \langle A_i, U\hat{X}^\top + \hat{X}U^\top\rangle^2 \tag{2.7b}$$

The proof to the above lemma is ubiquitous in many matrix optimization literature like [98, 94] and derived directly from multivariate calculus, thus omitted here for simplicity. Note that (2.5) and (2.6) are first and second-order conditions for general matrix objectives, and when constrained to the matrix sensing problem on which in dissertation is centered, we can further simplify ∇f and $\nabla^2 f$ as shown above, due to our knowledge of the sensing matrices, which serves as the cornerstone for most of the analysis done in this dissertation.

Global optimization landscape

Since this dissertation would focus on the characterization of the optimization landscape, and talk extensively about spurious solution, global solution, and strict saddles; we hope to give the readers a flavor of this kind of analysis before delving into the real results. Therefore, we adapt section 3.2 of [18] to offer a nice case study. One key objective of most of our works is prove that either 1) all SOPs of (1.3) is indeed the global solution or 2) all SOPs are close to the global solution. Consider this toy example of matrix sensing

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|xx^\top - M^*\|_F^2 \tag{2.8}$$

where we know $r = 1$, and that $\mathcal{A}(\cdot)$ is simply the identity operator. This corresponds to $\delta_{2r} = 0$, the most ideal case possible. We will proceed to show that all SOPs of (2.8) are global optima, and the rest of the critical points are either local maxima or strict saddle points.

We first characterize its critical points (FOPs), in which we require $\nabla f(x) = 0$, in this case means that

$$M^* x = \|x\|_2^2 x$$

which means that x is either 0 or eigenvectors of M^* scaled by square root of the respective eigenvalue, meaning

$$\hat{x} \in \{0, \pm\sqrt{\lambda_1} u_1, \pm\sqrt{\lambda_2} u_2, \dots, \pm\sqrt{\lambda_n} u_n\}$$

where λ_i, u_i are the i^{th} eigen-pair of M^*. Following the procedures given above, we hope to characterize these FOPs using their Hessian information. From elementary calculus, or if you so choose to apply Lemma 4, we get that

$$\nabla^2 f(x) = 2\|x\|_2^2 I_n + 4xx^\top - 2M^*$$

which means that

$$\nabla^2 f(\pm\sqrt{\lambda_i} u_i) = \sum_{j \neq i} 2(\lambda_i - \lambda_j) u_j u_j^\top + 4\lambda_i u_i u_i^\top$$

Therefore depending on the FOPs under consideration,

1. If $\hat{x} = \pm\sqrt{\lambda_1} u_1$, $\nabla^2 f \succ 0$ due to the above derivation, therefore they both are SOPs.

2. If $\hat{x} = \pm\sqrt{\lambda_i} u_i$ for $i \geq 2$, we can show that $\lambda_{\min}(\nabla^2 f) < 0$, therefore they are strict saddle points.

3. If $\hat{x} = 0$, $\nabla^2 f = -2M^* \preceq 0$, which is indeed negative semidefinite, thereby either a local maximum or a degenerate saddle point (if admits zero eigenvalues).

This shows that (2.8) only has two SOPs, which have the same value, therefore are all global solutions. Other critical points are either strict saddle points or local maxima, and will not cause great concern for gradient based optimization techniques [33]. The rest of this dissertation will revolve around this same problem with higher r and more complex $\mathcal{A}$, of course with more advanced theoretical tools than plain algebra.

Part I

The Under-Sampled Regime

Chapter 3

Convex Relaxation

3.1 Background and Related Work

It is widely known that the SDP formulation (1.2) can be used to solve the matrix sensing problem if the sensing matrices are sampled independently from a sub-Gaussian distribution and the number of measurements d is large enough [72, 73]. This is also a sufficient condition for the sensing matrices to satisfy the RIP condition with high probability, which is defined below:

The RIP constant δ_p represents how similar the linear operator $\mathcal{A}$ is to an isometry, and various upper bounds on δ_p have been proposed to serve as sufficient conditions for the exact recovery (meaning that one can recover the ground truth M^* by solving the SDP problem). A few notable ones include $\delta_{4r} < \sqrt{2} - 1$ in [13], $\delta_{5r} < 0.607, \delta_{3r} < 0.472$ in [60], and $\delta_{2r} < 1/2, \delta_{3r} < 1/3$ in [11]. On the other hand, when the sensing matrices are not sampled independently from a sub-Gaussian distribution or when the RIP condition is not met, the SDP formulation may still recover the ground truth matrix with a high probability. This is the case for MC problems for which RIP fails to hold while SDP works as long as entries of observation follow an independent Bernoulli model [15, 16].

However, recent works have shown that if we use the B-M method instead of the SDP approach, we can still recover the ground truth matrix via first-order methods under similar RIP or coherence assumptions in both the matrix sensing and matrix completion cases [24, 5, 65, 103, 106, 101, 93, 6, 27, 107, 95]. Namely, the state-of-the-art result states that as long as $\delta_{r_{\text{search}}+r} < 1/2$ for the matrix sensing problem, there exists no spurious local minima for an over-parametrized B-M formulation and the gradient descent algorithm can recover M^* exactly [95]. If we know the value of r, we can set r_{search} to r, making the B-M approach enjoy the same RIP guarantee as the SDP approach. Since the B-M approach enjoys far better scalability, it has become an increasingly popular tool for solving the matrix sensing problem.

Thus, it is important to investigate whether with the increased parametrization space (n^2 vs nr_{search}), SDP can enjoy better theoretical guarantees. This study is timely since specialized sparse SDP algorithms have become more efficient in recent years, making the SDP method more practical than before [99, 92, 91]. In this chapter, we show that the SDP approach is more powerful than the B-M method as far as the RIP measure is concerned.

3.2 Sharper RIP bound for SDP

Since the SDP method is more powerful than the B-M factorization for certain classes of MC and MS problems as shown in the previous section and since specialized SDP algorithms can solve large-scale MC and MS problems, it is useful to further study the SDP method through the lens of the well-known RIP notion. We will derive a strong lower bound δ_{lb} on the RIP constant δ to guarantee convergence to the ground truth solution by using a proof technique called the inexistence of incorrect solution [101]. We aim to find a linear measurement operator $\mathcal{A}$ with the smallest RIP constant such that the SDP formulation converges to a wrong solution. To do so, we need to solve the optimization problem

$$\begin{aligned} \min_{\delta,\mathcal{A}} \quad & \delta \\ \text{s.t.} \quad & \mathcal{A}(M) = \mathcal{A}(M^*) \\ & \operatorname{tr}(M) \leq \operatorname{tr}(M^*) \\ & \mathcal{A} \text{ satisfies the } \delta_{2r}\text{-RIP property,} \end{aligned} \tag{3.1}$$

where $M \neq M^*$. The condition $\operatorname{tr}(M) \leq \operatorname{tr}(M^*)$ guarantees that SDP cannot uniquely recover M^*. Checking the RIP constant for a linear measurement operator is proven to be NP-hard [77]. Therefore, it is difficult to solve the problem (3.1) analytically. To simplify the problem, we will introduce some notations. We use a matrix representation of the measurement operator $\mathcal{A}$ as follows:

$$\mathbf{A} = [\operatorname{vec}(\mathbf{A}_1), \operatorname{vec}(\mathbf{A}_2), \dots, \operatorname{vec}(\mathbf{A}_d)]^T \in \mathbb{R}^{d\times n^2}.$$

Then, $\mathbf{A}\operatorname{vec}(M) = \mathcal{A}(M)$ for every matrix $M \in \mathbb{R}^{n\times n}$. We define $\mathbf{H} = \mathbf{A}^T\mathbf{A}$, which is the matrix representation of the kernel operator $\mathcal{H} = \mathcal{A}^T\mathcal{A}$ to simplify the last constraint of the problem (3.1).

To derive a RIP bound, we consider the following optimization problem given M and M^*, where M is the global solution of (1.2) and M^* is the ground truth solution:

$$\begin{aligned} \min_{\delta,\mathbf{H}} \quad & \delta \\ \text{s.t.} \quad & \mathbf{e}^T\mathbf{H}\mathbf{e} = 0 \\ & \mathbf{H} \text{ is symmetric and satisfies the } \delta_{2r}\text{-RIP,} \end{aligned} \tag{3.2}$$

where

$$\mathbf{e} = \mathrm{vec}(M^* - M).$$

For this fixed M and M^*, we assume that $M \neq M^*$ and that $\mathrm{rank}(M^* - M) > 2r$, since if $\mathrm{rank}(M^* - M) \leq 2r$, the relation $M = M^*$ holds automatically by definition of δ_{2r}-RIP for any δ since it implies strong convexity. Denote the optimal value to (3.2) as $\delta(\mathbf{e})$, which is a function of $\mathbf{e}$. It is desirable to find

$$\delta^* := \min_{\mathbf{e}:\mathrm{tr}(M)\leq\mathrm{tr}(M^*)} \delta(\mathbf{e}).$$

By the logic of in-existence of counterexample, we know that if a problem $\mathbf{H} = \mathbf{A}^T\mathbf{A}$ has δ_{2r}-RIP with $\delta < \delta^*$, then the solution to (1.2) will be M^*, which is the ground truth solution. However, since the last constraint of (3.2) is non-convex, it is useful to replace it with a surrogate condition that allows solving the problem analytically. The following problem helps to achieve this goal:

$$\begin{aligned} \min_{\delta,\mathbf{H}} \quad & \delta \\ \text{s.t.} \quad & \mathbf{e}^T\mathbf{H}\mathbf{e} \leq 2\|\mathbf{e}_c\|^2 + 2(l-3)\delta\|\mathbf{e}_c\|^2 \\ & (1-\delta)\mathbf{I}_{n^2} \preceq \mathbf{H} \preceq (1+\delta)\mathbf{I}_{n^2}. \end{aligned} \tag{3.3}$$

Here, $l = \lceil n/r \rceil$ and we define $\{\mathbf{e}_i\}_{i=1}^l$ and $\mathbf{e}_c$ in the following fashion. First, consider the eigendecomposition of $M^* - M$ and assume that the eigenvalues are ordered in terms of their absolute values, namely, $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Let $\mathbf{u}_k$'s denote the corresponding orthonormal eigenvectors:

$$\mathrm{mat}_S(\mathbf{e}) = M^* - M = \sum_{k=1}^n \lambda_k \mathbf{u}_k \mathbf{u}_k^T.$$

Then, we define:

$$\mathbf{e}_i = \mathrm{vec}\left(\sum_{k=(i-1)*r+1}^{\min\{i*r,n\}} \lambda_k \mathbf{u}_k \mathbf{u}_k^T \right),$$

$\mathbf{e}_{2r} = \mathbf{e}_1 + \mathbf{e}_2$, and $\mathbf{e}_c = \sum_{i=3}^l \mathbf{e}_i$. The next proposition allows us to replace (3.2) with (3.3) because the optimal value of the (3.3), $\delta_{lb}(\mathbf{e})$, gives a lower bound on $\delta(\mathbf{e})$.

Proposition 1 *The optimal objective value of the problem* (3.3)*,* $\delta_{lb}(\mathbf{e})$*, is always less than or equal to the optimal objective value of the problem* (3.2)*, i.e.,* $\delta_{lb}(\mathbf{e}) \leq \delta(\mathbf{e})$*.*

The proof of this proposition is central to the construction of the sufficiency bound, which is based on using a convex program to serve as an estimate of the non-convex problem. After we extend the RIP_{2r} constraint in (3.3) to be RIP_n(thus making it convex), it is necessary to somehow preserve the information that the near isometric property of **H** should only apply to low-rank matrices. This is achieved by changing the first constraint so that **e** does not need to be completely in the null space of **H**. (3.3) approximately requires that **H** only maps a certain low-rank sub-manifold to 0. The full proof can be found in the Appendix. As a result of Proposition 1, it immediately follows that

$$\delta_{\mathrm{lb}} = \min_{\mathbf{e}:\mathrm{tr}(M)\leq \mathrm{tr}(M^*)} \delta_{\mathrm{lb}}(\mathbf{e}) \leq \delta^*.$$

In fact, we can obtain a lower bound on the value δ_{lb} by solving the problem (3.3) analytically. The following lemma quantifies a lower bound on δ_{lb}.

Lemma 5 *It holds that*

$$\delta_{lb} \geq \frac{2r}{n+(n-2r)(2l-5)}.$$

The best-known sufficiency bound presented in [11] is independent of n and r. This sufficiency lower bound on the RIP constant presented in Lemma 5 can be tighter than 1/2 depending on the size of the problem n and the rank of the ground truth matrix r. For instance, the SDP formulation converges to ground truth solution whenever RIP constant δ is close to 1 as $r \to n/2$. On the other hand, whenever r/n is ratio is small, e.g. rank-1 matrix sensing problem with large n, $\delta < 1/2$ is a stronger guarantee for recovery of the ground truth matrix. Combined with the 1/2 sufficiency bound that works for both the symmetric and asymmetric cases [11], we obtain the following result:

Theorem 1 *The global solution of the SDP formulation* (1.2) *will be the ground truth matrix M^* if the sensing matrix $\mathcal{A}$ satisfies the RIP condition with the RIP constant δ_{2r} satisfying the inequality:*

$$\delta_{2r} < \max\left\{1/2, \frac{2r}{n+(n-2r)(2l-5)}\right\},$$

where $l = \lceil n/r \rceil$.

Compared with the existing sufficiency RIP bounds, this new result has a striking advantage. The bound $\delta_{2r} < 1/2$ has already been proven to be the sharpest for the B-M formulation, which is independent of the search rank. In contrast, Theorem 1 shows that the RIP bound for SDP exceeds this bound and approaches 1 as the rank r increases.

In this section, we have shown that as opposed to the popular belief that B-M enjoys very similar RIP guarantees as the SDP approach, there are real benefits to switching to the SDP formulation, making it a more competitive option since specialized SDP solvers are becoming more efficient in recent years.

3.3 Summary

In this chapter, we conducted a comparison between two main approaches to the matrix completion and matrix sensing problems: a convex relaxation that gives an SDP formulation and the B-M factorization method. It is well-known that both of these methods enjoy mathematical guarantees for the recovery of the ground truth matrix whenever the RIP assumption is satisfied with a sufficiently small δ. We discovered classes of problems for which B-M factorization fails while the SDP recovers the ground truth matrix, namely in the regime in which $\delta_{2r} \geq 1/2$. The fact that specialized SDP algorithms are improved in recent years and can compete with simple first-order descent algorithms inspired us to investigate sharper bounds on sufficient conditions for the SDP formulation. We provided RIP bounds for the SDP formulation that depend on the rank of the solution and are automatically satisfied for high-rank problems, unlike the B-M method. As a result, we conclude that none of the methods outperforms the other one whenever the sufficiency guarantees are not met. The parameters of the problem, such as dimension, rank, and linear measurement operator, determine which solution method performs better. Consequently, it is prudent to apply both solution methods in case the RIP and incoherence are not satisfied.

3.A Missing Details of Chapter 3

Proof 2 (Proof of Proposition 1) *To prove Proposition 1, we study intermediary problem.*

$$\begin{aligned} \min_{\delta,\hat{\mathbf{H}}} \quad & \delta \\ s.t. \quad & \hat{\mathbf{e}}^T\hat{\mathbf{H}}\hat{\mathbf{e}} \leq (1+\delta)\|\mathbf{e}_c\|^2 + 2(l-3)\delta\|\mathbf{e}_c\|^2 \\ & (1-\delta)\mathbf{I}_{4r^2} \preceq \hat{\mathbf{H}} \preceq (1+\delta)\mathbf{I}_{4r^2}, \end{aligned} \tag{3.4}$$

where

$$\hat{\mathbf{e}} = \mathbf{P}^T\mathbf{e}, \qquad \mathbf{P} \in \mathbb{R}^{n^2\times 4r^2} = P \otimes P,$$

and $P \in \mathbb{R}^{n\times 2r}$ *is defined to be*

$$P = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_{2r},\end{bmatrix}$$

where $\mathbf{u}_i$*'s are orthonormal eigenvectors of* $M^* - M$ *so that* $P^TP = \mathbf{I}$*. Denote the optimal solution to* (3.4) *as* $\delta_P(\mathbf{e})$*. Then, the following two lemmas will suffice to prove Proposition 1.*

Lemma 6 *Given a fixed vector* $\mathbf{e} \in \mathbb{R}^{n^2}$*, we have*

$$\delta_P(\mathbf{e}) \leq \delta(\mathbf{e}). \tag{3.5}$$

Lemma 7 *Given a fixed vector* $\mathbf{e} \in \mathbb{R}^{n^2}$*, we have*

$$\delta_{lb}(\mathbf{e}) \leq \delta_P(\mathbf{e}). \tag{3.6}$$

Proof 3 (Proof of Lemma 6) *It suffices to show that for any feasible pair* $(\delta, \bar{\mathbf{H}})$ *of* (3.2)*, we can construct a feasible solution* $(\delta, \hat{\mathbf{H}})$ *to* (3.4) *characterized as below*

$$\delta = \delta, \qquad \hat{\mathbf{H}} = \mathbf{P}^T\bar{\mathbf{H}}\mathbf{P},$$

which directly proves the lemma. We can verify the feasibility of $(\delta, \hat{\mathbf{H}})$ *as follows. The feasibility of the first constraint is certified by the following argument:*

$$\hat{\mathbf{e}}^T\hat{\mathbf{H}}\hat{\mathbf{e}} = \mathbf{e}^T\mathbf{P}\mathbf{P}^T\bar{\mathbf{H}}\mathbf{P}\mathbf{P}^T\mathbf{e},$$

By the definition of $\mathbf{P}$*, one can write*

$$\mathbf{P}\mathbf{P}^T\mathbf{e} = (PP^T \otimes PP^T)\mathbf{e} = \text{vec}(PP^T(M^*-M)PP^T) = \mathbf{e}_1 + \mathbf{e}_2 = \mathbf{e}_{2r},$$

Since $\mathbf{e}^T\bar{\mathbf{H}}\mathbf{e} = 0$ *and* $\bar{\mathbf{H}}$ *is symmetric,* $\bar{\mathbf{H}}$ *admits a factorization* $\bar{\mathbf{H}} = \bar{\mathbf{A}}^T\bar{\mathbf{A}}$*, making* $\bar{\mathbf{A}}\mathbf{e} = 0$*. Also, we know that* $\mathbf{e} = \mathbf{e}_{2r} + \mathbf{e}_c$*, meaning that*

$$\bar{\mathbf{A}}\mathbf{e}_{2r} = -\bar{\mathbf{A}}\mathbf{e}_c.$$

Therefore,

$$\hat{\mathbf{e}}^T\hat{\mathbf{H}}\hat{\mathbf{e}} = \mathbf{e}_{2r}^T\bar{\mathbf{H}}\mathbf{e}_{2r} = \mathbf{e}_c^T\bar{\mathbf{H}}\mathbf{e}_c = (\sum_{i=3}^{l}\mathbf{e}_i)^T\bar{\mathbf{H}}(\sum_{i=3}^{l}\mathbf{e}_i).$$

Since $\bar{\mathbf{H}}$ *satisfies* δ_{2r}*-RIP, for every* (i,j) *such that* $i \neq j$*, we have:*

$$(\mathbf{e}_i+\mathbf{e}_j)^T\bar{\mathbf{H}}(\mathbf{e}_i+\mathbf{e}_j) \leq (1+\delta)\|\mathbf{e}_i+\mathbf{e}_j\|^2 = (1+\delta)(\|\mathbf{e}_i\|^2+\|\mathbf{e}_j\|^2), \tag{3.7}$$

where the last equality follows from the facts that $\mathbf{e}_i^T\mathbf{e}_j = 0$ *and*

$$(\mathbf{e}_i+\mathbf{e}_j)^T\bar{\mathbf{H}}(\mathbf{e}_i+\mathbf{e}_j) = \mathbf{e}_i^T\bar{\mathbf{H}}\mathbf{e}_i + 2\mathbf{e}_i^T\bar{\mathbf{H}}\mathbf{e}_j + \mathbf{e}_j^T\bar{\mathbf{H}}\mathbf{e}_j \geq 2\mathbf{e}_i^T\bar{\mathbf{H}}\mathbf{e}_j + (1-\delta)(\|\mathbf{e}_i\|^2+\|\mathbf{e}_j\|^2). \tag{3.8}$$

Combining (3.7) *and* (3.8) *yields that*

$$\mathbf{e}_i^T\bar{\mathbf{H}}\mathbf{e}_j \leq \delta(\|\mathbf{e}_i\|^2+\|\mathbf{e}_j\|^2) \qquad \forall i \neq j.$$

Therefore,

$$\begin{aligned}\hat{\mathbf{e}}^T\hat{\mathbf{H}}\hat{\mathbf{e}} = (\sum_{i=3}^{l}\mathbf{e}_i)^T\bar{\mathbf{H}}(\sum_{i=3}^{l}\mathbf{e}_i) &\leq (1+\delta)(\sum_{i=3}^{l}\|\mathbf{e}_i\|^2) + 2\delta(l-3)(\sum_{i=3}^{l}\|\mathbf{e}_i\|^2)\\ &= (1+\delta)\|\mathbf{e}_c\|^2 + 2\delta(l-3)\|\mathbf{e}_c\|^2.\end{aligned}$$

The above inequality directly verifies the satisfaction of the first constraint. For the second constraint, consider an arbitrary vector $\tilde{\mathbf{e}} \in \mathbb{R}^{4r^2}$*. Then,*

$$\begin{aligned}\tilde{\mathbf{e}}^T\hat{\mathbf{H}}\tilde{\mathbf{e}} = \tilde{\mathbf{e}}^T\mathbf{P}^T\bar{\mathbf{H}}\mathbf{P}\tilde{\mathbf{e}} &= \tilde{\mathbf{e}}^T(P^T\otimes P^T)\bar{\mathbf{H}}(P\otimes P)\tilde{\mathbf{e}}\\ &= \operatorname{vec}(P\operatorname{mat}(\tilde{\mathbf{e}})P^T)^T\bar{\mathbf{H}}\operatorname{vec}(P\operatorname{mat}(\tilde{\mathbf{e}})P^T).\end{aligned}$$

By orthogonal projection, we know that $P\operatorname{mat}(\tilde{\mathbf{e}})P^T \in \mathbb{R}^{n\times n}$ *has rank* $2r$*. Therefore, the following holds by the* δ_{2r}*-RIP property of* $\bar{\mathbf{H}}$*:*

$$(1-\delta)\|P\operatorname{mat}(\tilde{\mathbf{e}})P^T\|_F^2 \leq \operatorname{vec}(P\operatorname{mat}(\tilde{\mathbf{e}})P^T)^T\bar{\mathbf{H}}\operatorname{vec}(P\operatorname{mat}(\tilde{\mathbf{e}})P^T) \leq (1+\delta)\|P\operatorname{mat}(\tilde{\mathbf{e}})P^T\|_F^2 \tag{3.9}$$

and since

$$\begin{aligned}\|P\operatorname{mat}(\tilde{\mathbf{e}})P^T\|_F^2 &= \operatorname{tr}(P\operatorname{mat}(\tilde{\mathbf{e}})^TP^TP\operatorname{mat}(\tilde{\mathbf{e}})P^T)\\ &= \operatorname{tr}(P\operatorname{mat}(\tilde{\mathbf{e}})^T\operatorname{mat}(\tilde{\mathbf{e}})P^T)\\ &= \operatorname{tr}(P^TP\operatorname{mat}(\tilde{\mathbf{e}})^T\operatorname{mat}(\tilde{\mathbf{e}}))\\ &= \operatorname{tr}(\operatorname{mat}(\tilde{\mathbf{e}})^T\operatorname{mat}(\tilde{\mathbf{e}}))\\ &= \|\tilde{\mathbf{e}}\|_2^2,\end{aligned}$$

(3.9) *automatically implies the satisfaction of the second constraint.*

Proof 4 (Proof of Lemma 7) *It suffices to show that for any feasible pair* $(\delta, \hat{\mathbf{H}})$ *of* (3.4)*, we can construct a feasible solution* $(\delta, \mathbf{H})$ *to* (3.3) *characterized as*

$$\delta = \delta, \qquad \mathbf{H} = \mathbf{P}\hat{\mathbf{H}}\mathbf{P}^T + (1-\delta)(\mathbf{I}_{n^2} - \mathbf{P}\mathbf{P}^T).$$

To prove the lemma, it is enough to verify that the above pair $(\delta, \mathbf{H})$ *is feasible to* (3.3)*. We first verify the second constraint. Given an arbitrary vector* $\mathbf{e} \in \mathbb{R}^{n^2}$*, we have that*

$$\mathbf{e}^T\mathbf{H}\mathbf{e} = \mathbf{e}^T\mathbf{P}\hat{\mathbf{H}}\mathbf{P}^T\mathbf{e} + (1-\delta)\left[\mathbf{e}^T\mathbf{e} - \mathbf{e}^T\mathbf{P}\mathbf{P}^T\mathbf{e}\right]$$

and defining $\tilde{\mathbf{e}} := \mathbf{P}^T\mathbf{e} \in \mathbb{R}^{4r^2}$*, we obtain:*

$$\mathbf{e}^T\mathbf{P}\hat{\mathbf{H}}\mathbf{P}^T\mathbf{e} + (1-\delta)\left[\mathbf{e}^T\mathbf{e} - \mathbf{e}^T\mathbf{P}\mathbf{P}^T\mathbf{e}\right] \geq (1-\delta)\|\tilde{\mathbf{e}}\|_2^2 + (1-\delta)[\|\mathbf{e}\|_2^2 - \|\tilde{\mathbf{e}}\|_2^2] = (1-\delta)\|\mathbf{e}\|_2^2.$$

Also, since $\|\tilde{\mathbf{e}}\|_2^2 \leq \|\mathbf{e}\|_2^2$ *and* P *is a projection matrix, one can write:*

$$(1+\delta)[\|\mathbf{e}\|_2^2 - \|\tilde{\mathbf{e}}\|_2^2] \geq (1-\delta)[\|\mathbf{e}\|_2^2 - \|\tilde{\mathbf{e}}\|_2^2],$$

which further implies that

$$\mathbf{e}^T\mathbf{P}\hat{\mathbf{H}}\mathbf{P}^T\mathbf{e} + (1-\delta)\left[\mathbf{e}^T\mathbf{e} - \mathbf{e}^T\mathbf{P}\mathbf{P}^T\mathbf{e}\right] \leq (1+\delta)\|\tilde{\mathbf{e}}\|_2^2 + (1+\delta)[\|\mathbf{e}\|_2^2 - \|\tilde{\mathbf{e}}\|_2^2] = (1+\delta)\|\mathbf{e}\|_2^2.$$

Combining the above equations, we recover the second constraint of (3.3)*:*

$$(1-\delta)\|\mathbf{e}\|_2^2 \leq e^T\mathbf{H}\mathbf{e} \leq (1+\delta)\|\mathbf{e}\|_2^2.$$

To study the first constraint, we have that

$$\begin{aligned} e^T\mathbf{H}e &= \hat{\mathbf{e}}^T\hat{\mathbf{H}}\hat{\mathbf{e}} + (1-\delta)\left[\|\mathbf{e}\|_2^2 - \|\hat{\mathbf{e}}\|_2^2\right] \\ &\leq (1+\delta)\|\mathbf{e}_c\|_2^2 + 2(l-3)\delta\|\mathbf{e}_c\|_2^2 + (1-\delta)\|\mathbf{e}_c\|_2^2 \\ &= 2\|\mathbf{e}_c\|_2^2 + 2(l-3)\delta\|\mathbf{e}_c\|_2^2. \end{aligned}$$

Note that $\|\mathbf{e}\|_2^2 - \|\hat{\mathbf{e}}\|_2^2 = \|\mathbf{e}_c\|_2^2$ *due to*

$$\|\mathbf{e}\|_2^2 = \sum_{i=1}^{n} \lambda_i^2, \qquad \|\hat{\mathbf{e}}\|_2^2 = \sum_{i=1}^{2r} \lambda_i^2.$$

The proof of Proposition 1 follows directly from combining Lemma 6 and 7.

Proof 5 (Proof of Lemma 5) *We aim to solve* (3.3) *analytically to obtain a sufficient RIP bound for problem* (1.2)*. This amounts to deriving a closed-form expression for* $\delta_{lb}(\mathbf{e})$*. We consider a simpler problem to solve* (3.3)*:*

$$\begin{aligned} \max_{\eta, \tilde{\mathbf{H}}} \quad & \eta \\ \text{s.t.} \quad & \mathbf{e}^T\tilde{\mathbf{H}}\mathbf{e} \leq \frac{1-\eta}{2}c^2 + \frac{1+\eta}{2}d^2 \\ & \eta I_{n^2} \preceq \tilde{\mathbf{H}} \preceq I_{n^2} \end{aligned} \tag{3.10}$$

with $c^2 = 2(l-3)\|\mathbf{e}_c\|_2^2$ *and* $d^2 = 2\|\mathbf{e}_c\|_2^2$. *Given any feasible solution* $(\delta, \mathbf{H})$ *to* (3.3), *the tuple*

$$\Big(\frac{1-\delta}{1+\delta}, \frac{1}{1+\delta}\mathbf{H}\Big)$$

is a feasible solution to problem (3.10). *Therefore, if we denote the optimal value of* (3.10) *as* $\eta(\mathbf{e})$, *then it holds that*

$$\eta(\mathbf{e}) \geq \frac{1-\delta_{lb}(\mathbf{e})}{1+\delta_{lb}(\mathbf{e})} \implies \delta_{lb}(\mathbf{e}) \geq \frac{1-\eta(\mathbf{e})}{1+\eta(\mathbf{e})}. \tag{3.11}$$

We use the dual problem to solve for $\eta(\mathbf{e})$:

$$\begin{aligned}
\min_{\mathbf{U}_1,\mathbf{U}_2,\gamma} \quad & \operatorname{tr}(\mathbf{U}_2) + \frac{\gamma}{2}(c^2+d^2) \\
\text{s. t.} \quad & \operatorname{tr}(\mathbf{U}_1) + \frac{\gamma}{2}(c^2-d^2) = 1 \\
& \gamma\mathbf{e}\mathbf{e}^T = \mathbf{U}_1 - \mathbf{U}_2, \quad \mathbf{U}_1, \mathbf{U}_2 \succeq 0,\ \gamma \geq 0.
\end{aligned} \tag{3.12}$$

Since Slater's condition holds for the convex program (3.10), *the optimal solution to* (3.12) *is equivalent to that of* (3.10), *which is* $\eta(\mathbf{e})$. *Using a Lagrangian argument,* $\eta(\mathbf{e})$ *can be solved as follows:*

$$\begin{aligned}
\eta(\mathbf{e}) &= \max_{\beta\in\mathbb{R}} \min_{\gamma\geq 0} \left\{ \beta(1 - \frac{\gamma}{2}(c^2-d^2)) + \gamma\frac{c^2+d^2}{2} + \min_{\substack{\mathbf{U}_1\succeq 0 \\ \mathbf{U}_1 - \gamma\mathbf{e}\mathbf{e}^T \succeq 0}} \left[\operatorname{tr}(\mathbf{U}_1 - \gamma\mathbf{e}\mathbf{e}^T) - \beta\operatorname{tr}(\mathbf{U}_1)\right] \right\} \\
&= \max_{\beta\leq 1} \min_{\gamma\geq 0} \left\{ \beta(1 - \frac{\gamma}{2}(c^2-d^2)) + \gamma\frac{c^2+d^2}{2} + \min_{\substack{\mathbf{U}_1\succeq 0 \\ \mathbf{U}_1 - \gamma\mathbf{e}\mathbf{e}^T \succeq 0}} \left[(1-\beta)\operatorname{tr}(\mathbf{U}_1) - \gamma\|\mathbf{e}\|_2^2\right] \right\} \\
&= \max_{\beta\leq 1} \min_{\gamma\geq 0} \left\{ \beta(1 - \frac{\gamma}{2}(c^2-d^2)) + \gamma\frac{c^2+d^2}{2} + \gamma(1-\beta)\|\mathbf{e}\|_2^2 - \gamma\|\mathbf{e}\|_2^2 \right\} \\
&= \max_{\beta\leq 1} \left\{ \beta + \min_{\gamma\geq 0}\left[\gamma(\frac{c^2+d^2}{2} - \beta(\frac{c^2-d^2}{2} + \|\mathbf{e}\|_2^2))\right] \right\} \\
&= \max_{\beta\leq 1} \left\{ \beta : \frac{c^2+d^2}{2} - \beta(\frac{c^2-d^2}{2} + \|\mathbf{e}\|_2^2) \geq 0 \right\} \\
&= \min\left\{ 1, \frac{c^2+d^2}{2\|\mathbf{e}\|_2^2 + c^2 - d^2} \right\},
\end{aligned}$$

where the first equality uses the Lagrangian argument by introducing the Lagrange multiplier β, *and the second equality constraints* $\beta \leq 1$ *since* $(1-\beta)\operatorname{tr}(\mathbf{U}_1)$ *will be unbounded otherwise. The third equality results from the obvious choice of* $\mathbf{U}_1 = \gamma\mathbf{e}\mathbf{e}^T$ *given that* $(1-\beta)$ *is nonnegative. The fifth equality results from the choice of* $\gamma = 0$ *constrained to the requirement that its coefficient must be nonnegative.*

Substituting $\eta(\mathbf{e}) = 1$ into (7.31) results in the trivial lower bound $\delta_{lb}(\mathbf{e})$ of 0, which means that the lower bound indeed will not be negative. Hence, we will focus on $\frac{c^2+d^2}{2\|\mathbf{e}\|_2^2+c^2-d^2}$ from now on. We know from (7.31) that in order to obtain a lower bound on δ_{lb}, we need to derive an upper bound on $\eta(\mathbf{e})$. Note that

$$\frac{c^2+d^2}{2\|\mathbf{e}\|_2^2+c^2-d^2} = \frac{2(l-2)\|\mathbf{e}_c\|_2^2}{2\|\mathbf{e}\|_2^2+2(l-4)\|\mathbf{e}_c\|_2^2} = \frac{(l-2)\|\mathbf{e}_c\|_2^2}{\|\mathbf{e}_{2r}\|_2^2+(l-3)\|\mathbf{e}_c\|_2^2}. \tag{3.13}$$

The last equality follows from the relations

$$\|\mathbf{e}\|_2^2 = \|\mathbf{e}_c+\mathbf{e}_{2r}\|_2^2 = \|\mathbf{e}_{2r}\|_2^2+\|\mathbf{e}_c\|_2^2.$$

If we fix $\|\mathbf{e}_{2r}\|_2^2$, then we can maximize (3.13) with respect to $\|\mathbf{e}_c\|_2^2$ first. In this case, taking the derivative of (3.13) yields that

$$\frac{\partial}{\partial\|\mathbf{e}_c\|_2}\left(\frac{(l-2)\|\mathbf{e}_c\|_2^2}{\|\mathbf{e}_{2r}\|_2^2+(l-3)\|\mathbf{e}_c\|_2^2}\right) = 2\frac{(l-2)\|\mathbf{e}_c\|_2\|\mathbf{e}_{2r}\|_2^2}{(\|\mathbf{e}_{2r}\|_2^2+(l-3)\|\mathbf{e}_c\|_2^2)^2} \geq 0.$$

Therefore, (3.13) is maximized when $\|\mathbf{e}_c\|_2^2$ is set to be as large as possible. Before we derive the maximum value of $\|\mathbf{e}_c\|_2^2$ in terms of $\|\mathbf{e}_1\|_2^2$ and $\|\mathbf{e}_2\|_2^2$, we introduce one key lemma.

Lemma 8 *Consider two PSD matrices M and M^* such that $\operatorname{tr}(M) \leq \operatorname{tr}(M^*)$ and $\operatorname{rank}(M^*) = r$. Then,*

$$\sigma_{(1)}(M^*-M)+\cdots+\sigma_{(r)}(M^*-M) \geq \sigma_{(r+1)}(M^*-M)+\cdots\sigma_{(n)}(M^*-M), \tag{3.14}$$

where σ_i denote the i-th largest singular value of the matrix M^*-M.

Proof 6 *For each matrix $\mathbf{A}$, we denote the i^{th} eigenvalue as $\lambda_{(i)}(\cdot)$, meaning that*

$$\lambda_{(1)}(\mathbf{A}) \geq \lambda_{(2)}(\mathbf{A}) \geq \cdots \geq \lambda_{(n)}(\mathbf{A}).$$

By Weyl's inequality, we know that

$$\lambda_{(i+j-1)}(M^*-M) \leq \lambda_{(i)}(M^*)+\lambda_{(j)}(-M).$$

Hence,

$$\lambda_{(r+1)}(M^*-M) \leq \lambda_{(r+1)}(M^*)+\lambda_{(1)}(-M) \leq 0$$

since M^* is of rank-r and $M \succeq 0$. Therefore, we know that M^*-M has at most r positive eigenvalues. Also, since $\operatorname{tr}(M^*-M) \geq 0$, it holds that

$$\lambda_{(1)}(M^*-M)+\cdots+\lambda_{(r)}(M^*-M) \geq -\lambda_{(r+1)}(M^*-M)-\cdots-\lambda_{(n)}(M^*-M),$$

which implies that

$$|\lambda_{(1)}(M^*-M)|+\cdots+|\lambda_{(r)}(M^*-M)| \geq |\lambda_{(r+1)}(M^*-M)|+\cdots+|\lambda_{(n)}(M^*-M)| \quad (3.15)$$

since $\lambda_{(k)}(M^* - M) \leq 0$ *for all* $k > r$. *According to the definition, we have*

$$\begin{aligned} &|\lambda_1(M^* - M)| + \cdots + |\lambda_r(M^* - M)| \geq |\lambda_{(1)}(M^* - M)| + \cdots + |\lambda_{(r)}(M^* - M)|, \\ &|\lambda_{(r+1)}(M^* - M)| + \cdots + |\lambda_{(n)}(M^* - M)| \geq |\lambda_{r+1}(M^* - M)| + \dots |\lambda_n(M^* - M)| \end{aligned} \quad (3.16)$$

since $\lambda_1(M^* - M), \dots, \lambda_n(M^* - M)$ *are ordered with respect to their absolute values. As per the main text, we abbreviate* $\lambda_i(M^* - M)$ *as* λ_i *for the sake of brevity for any* $i \in [n]$. *Combining* (3.15) *with* (3.16) *proves the original lemma.*

Denote $S_1 := \sum_{i=1}^{r} |\lambda_i|$ *and* $S_2 := \sum_{i=r+1}^{n} |\lambda_i|$. *Given* S_2, *since* $|\lambda_i| \leq |\lambda_r|$ *as long as* $i > r$, *we know that* $\sum_{i=r+1}^{n} \lambda_i^2$ *is maximized when every absolute value is chosen to be as large as possible, namely*

$$|\lambda_{r+1}|, \dots, |\lambda_{r+\lfloor S_2/|\lambda_r|\rfloor}| = |\lambda_r|, |\lambda_{r+\lceil S_2/|\lambda_r|\rceil}| = S_2 - \lfloor S_2/|\lambda_r|\rfloor|\lambda_r| := \tilde{\lambda} \leq |\lambda_r|.$$

Therefore,

$$\sum_{i=r+1}^{n} \lambda_i^2 \leq \lfloor S_2/|\lambda_r|\rfloor \lambda_r^2 + \tilde{\lambda}^2 \leq \lfloor S_2/|\lambda_r|\rfloor \lambda_r^2 + \frac{\tilde{\lambda}}{|\lambda_r|}\lambda_r^2 = \frac{S_2}{|\lambda_r|}\lambda_r^2.$$

As a result,

$$\frac{S_2}{|\lambda_r|}\lambda_r^2 = S_2|\lambda_r| \leq S_1|\lambda_r| \leq S_1\frac{S_1}{r} \leq \frac{S_1^2}{r} \leq \frac{r\|\mathbf{e}_1\|_2^2}{r} = \|\mathbf{e}_1\|_2^2,$$

where the last inequality follows from Cauchy-Schwartz. Combining the above 2 inequalities, we obtain

$$\sum_{i=r+1}^{n} \lambda_i^2 \leq \|\mathbf{e}_1\|_2^2. \quad (3.17)$$

Furthermore, since $|\lambda_{r+1}| \geq \cdots \geq |\lambda_n|$, *one can write:*

$$\|\mathbf{e}_c\|_2^2 = \sum_{i=2r+1}^{n} \lambda_i^2 \leq \frac{n-2r}{n-r} \sum_{i=r+1}^{n} \lambda_i^2$$

with equality holding if and only if $|\lambda_{r+1}| = \cdots = |\lambda_n|$. *Combined with* (3.17), *we obtain*

$$\|\mathbf{e}_c\|_2^2 \leq \frac{n-2r}{n-r}\|\mathbf{e}_1\|_2^2. \quad (3.18)$$

Consequently,

$$\|\mathbf{e}_c\|_2^2 \leq \frac{n-2r}{n-r}(\|\mathbf{e}_2\|_2^2 + \|\mathbf{e}_c\|_2^2) \implies \|\mathbf{e}_2\|_2^2 \geq \frac{r}{n-2r}\|\mathbf{e}_c\|_2^2. \tag{3.19}$$

It results from (3.18) *and* (3.19) *that*

$$\begin{aligned}\max_{\mathbf{e}:\mathrm{tr}(M)\leq\mathrm{tr}(M^*)} \eta(\mathbf{e}) &= \max_{\mathbf{e}:\mathrm{tr}(M)\leq\mathrm{tr}(M^*)} \frac{(l-2)\|\mathbf{e}_c\|_2^2}{\|\mathbf{e}_{2r}\|_2^2 + (l-3)\|\mathbf{e}_c\|_2^2} \\ &\leq \frac{(l-2)\|\mathbf{e}_c\|_2^2}{\|\mathbf{e}_1\|_2^2 + \frac{r}{n-2r}\|\mathbf{e}_c\|_2^2 + (l-3)\|\mathbf{e}_c\|_2^2} \\ &\leq \frac{(l-2)\frac{n-2r}{n-r}}{1 + (\frac{r}{n-2r} + l - 3)\frac{n-2r}{n-r}} \\ &= \frac{(l-2)(n-2r)}{n + (n-2r)(l-3)}.\end{aligned}$$

Thus,

$$\delta_{lb} \geq \max_{\mathbf{e}:\mathrm{tr}(M)\leq\mathrm{tr}(M^*)} \frac{1-\delta(\mathbf{e})}{1+\delta(\mathbf{e})} = \frac{2r}{n + (n-2r)(2l-5)}.$$

Chapter 4

Lifted Tensor Model

4.1 Background and Related Work

As we have extensively talked in Section 1.3, the solving of our main question of interest (1.3) depends on the RIP constant, which we restate here for clarity

$$\min_{X \in \mathbb{R}^{n \times r_{\text{search}}}} h(X) := \frac{1}{2}\|\mathcal{A}(XX^T) - b\|^2$$

The line of work [101, 6, 94] has shown that if (1.3) satisfies the RIP condition with $\delta_{2r} < 1/2$, every local minimizer $\hat{X}$ of (1.3) will satisfy the relation $\hat{X}\hat{X}^\top = M^*$, precisely in the noiseless scenario. It has also been proven that $\delta_{2r} < 1/2$ is a sharp bound, meaning that there are counterexamples such that $\hat{X}\hat{X}^\top \neq M^*$ once $\delta_{2r} \geq 1/2$. This also falls in line with a prior result that $\delta_{2r} < 1/2$ is sufficient for recovering M^* using specialized methods directly applied to our SDP formulation (1.2)[72, 14, 11].

The bound $\delta_{2r} < 1/2$ is sharp, and RIP conditions are difficult to satisfy and verify except for isometric Gaussian observations. In many applications, such as power system analysis, the RIP constant does not exist or is above 0.99 [100]. Yet, it is highly desirable to transfer the scalability benefits of the BM factorization approach to these practical cases as well. Hence, it is essential to investigate how to handle problems that do not satisfy the RIP property with a constant smaller than 1/2, using BM-type techniques. Towards this end, an active line of research has studied the relationship between the complexity of recovering the global optimum and the degree of (over-) parametrization in (1.3) [96, 97, 44], and the results are promising.

The current idea of over-parametrization in matrix sensing consists of enlarging the search space of X from $\mathbb{R}^{n \times r}$ to $\mathbb{R}^{n \times r_{\text{search}}}$, where $r_{\text{search}} \in [r, n)$. The above-mentioned papers have shown that as r_{search} increases, stronger guarantees for the recovery of M^* can be obtained (although it requires stricter assumptions). One of the main results in this area will be stated below.

Theorem 2 (Theorem 1.1 of [97]) *Assume that* (1.3) *satisfies the* (L_s, n)*-RSS (Restricted Strong Smoothness) and* (α_s, n)*-RSC (Restricted Smooth Convexity) properties. If*

$$r_{search} > \frac{1}{4}\left(\frac{L_s}{\alpha_s} - 1\right)^2 r, \quad r \leq r_{search} < n \tag{4.1}$$

then every second-order point (SOP) $\hat{X} \in \mathbb{R}^{n \times r_{search}}$ *of* (1.3) *satisfies that* $\hat{X}\hat{X}^\top = M^*$.

Note that $\hat{X}$ is an SOP if it satisfies the first-order and second-order necessary optimality conditions. The above theorem replaces the RIP condition with the similar conditions of RSS and RSC, which will be formally defined in the next section. The power of this theorem is in dealing with the scenario where $\delta_{2r} \geq 1/2$, by selecting a search rank $r_{\text{search}} > r$.

Despite the superiority of setting $r_{\text{search}} > r$ over setting $r_{\text{search}} = r$ under the setting of (1.3), the power of over-parametrization is limited. The reason is that r_{search} cannot be greater than n and therefore it is impossible to satisfy the condition (4.1) in practical cases where L_s/α_s is large. This calls for a new framework that accommodates an arbitrarily large degree of parametrization (as opposed to $r_{\text{search}} < n$), which would be effective in the regime of high L_s/α_s values. In this paper, we address this problem by proposing a tensor-based framework and analyzing its optimization landscape.

Overall, over-parametrization is a powerful idea since the inclusion of extra variables reshapes the landscape of the problem. Outside the realm of matrix sensing, the idea of constructing an infinite hierarchy of non-convex problems of increasing dimensions has been applied to the Tensor PCA problem [83]. The empirical evidence of deep learning practice shows the advantage of using overparametrized models for both convergence properties during training [64, 109, 22, 2] and generalization performance of the trained model [1, 62, 59, 4]. Practitioners also design their own hierarchy of machine learning models, to satisfy the scaling laws [37, 29, 58]. The cornerstone idea on the theoretical side of this field comes from the development of a hierarchy of convex problems, called the Sum-of-Squares hierarchy.

4.1.1 Sum-of-Squares Optimization

One of the most prominent over-parametrization frameworks for polynomial optimization is the framework of Sum-of-Squares (SOS) hierarchy of optimization problems [66, 43]. SOS optimization is essentially an optimization framework that leverages deep results in algebraic geometry to construct a hierarchy of convex problems of increasing qualities, solving each of which obtains a lower-bound certificate on the minimum value of the polynomial optimization problem of interest. Since (1.3) is also a polynomial optimization problem, SOS can be applied to handle the problem through a highly parametrized setting. Moreover, instead of using the usual SOS framework that finds a sequence of lower bounds on the optimal value of (1.3), we could use its dual problem, since the minimum value of (1.3) is 0 by construction. To construct the dual SOS problem, define $\kappa \geq 1$ to be an integer such that 2κ is equal to or larger than the maximum degree of $f(v)$ in (1.3), where $v := \mathrm{vec}(X)$. Here for simplicity please assume $r = 1$. Furthermore, define $[v]_\kappa \in \mathbb{R}^s$ to be a vector containing the standard monomials of v up to degree κ, with $s := \binom{n+\kappa}{\kappa}$. We then build the moment matrix $D := [v]_\kappa [v]_\kappa^\top$ with its entries being all standard monomials up to degree 2κ. As a result, it is possible to rewrite $f(v)$ (i.e., $f(X)$) as a linear function of D, namely

$$f(v) = \langle F, D \rangle$$

for some constant matrix $F \in \mathbb{R}^{s \times s}$. Therefore, optimizing $\langle F, D \rangle$ is equivalent to optimizing $f(v)$ given that D is rank-1 and positive-semidefinite. However, the rank-1 constraint is non-convex and its elimination leads to the dual SOS problem with the following form:

$$\min_{D \in \mathbb{S}^s} \langle F, D \rangle \text{ s.t. } \mathcal{L}(D) = 0, D \succeq 0 \tag{4.2}$$

The linear operator $\mathcal{L}$ captures the so-called consistency constraints, as some entries in D may be identical due to being the outer product of monomial vectors. For example, if $n = 2, \kappa = 2$, we have

$$[v]_\kappa = [1, v_1, v_2, v_1^2, v_1 v_2, v_2^2]^\top$$

meaning that $D_{15} = D_{23} = v_1 v_2$, $D_{14} = D_{22} = v_1^2$, $D_{34} = D_{25} = v_1^2 v_2$, $D_{26} = D_{35} = v_1 v_2^2$, and so on. The dual SOS problem (4.2) has some nice properties: it is convex and its optimal value asymptotically reaches that of (1.3) as κ grows to infinity (under generic conditions), which enables solving the non-convex (1.3) with an arbitrary accuracy [43]. However, the problem (4.2) also presents daunting challenges.

First, it has poor scalability properties because it requires solving costly SDP problems. The idea behind this paper is related to applying the BM factorization to (1.3) (without dealing with SDPs) via a lifting technique similar to (4.2). Currently, there is no guarantee that local minimizers of the BM formulation will translate to the minimizer of the convex problem (4.2). The state-of-the-art result regarding the BM factorization states that this correspondence can be established only when $r(r + 1)/2 \geq m$, where m is the number of linear constraints [8]. In matrix sensing, since r is small and m is large, this result cannot be applied.

Second, it is difficult to gauge how large κ needs to be in order for the convex relaxation to be exact, meaning that one may need to use significant computational resources to solve an instance of (4.2) corresponding to some value of κ, only to discover that its solution does not provide useful information about the optimal solution of the original problem, promoting to repeat the process for a larger value of

κ. This also prevents the practical application of SOS as it is common to miscalculate in advance how computationally challenging it can be to solve (1.3) via the SOS framework.

4.1.2 Related Works

Algorithm regularization in over-parametrized matrix sensing. [48, 108] prove that the convergence to global solution via GD is agnostic of r_{search}, in that it only depends on initialization scale, step-size, and RIP property. [51] demonstrates the same effect for an l_1 norm, and further showed that a small initialization nullifies the effect of over-parametrization. Besides these works, [76] refined this analysis, showing that via a sufficiently small initialization, the GD trajectory will make the solution implicitly penalize towards rank-r matrices after a small number of steps. [34] took it even further by showing that the GD trajectory will first make the matrix rank-1, rank-2, all the way to rank-r, in a sequential way, thereby resembling incremental learning.

Implicit bias in tensor learning. The line of work [70, 71, 25] demonstrates that for a class of tensor factorization problems, as long as the initialization scale is small, the learned tensor via GD will be approximately rank-1 after an appropriate number of steps. Our paper differs from this line of work in three meaningful ways: 1) The problem considered in those works are optimization problems over vectors, not tensors, and therefore the goal is to learn the structure of a known tensor, rather than learning a tensor itself; 2) Our proof relies directly on tensor algebra instead of adopting a dynamical systems perspective, providing deeper insights into tensor training dynamics while dispensing with the impractical assumption of an infinitesimal step-size.

4.2 The Tensor Framework

In this paper, we build upon some of the core ideas of SOS optimization in order to construct a new framework for over-parametrization that addresses the current issues with (4.2). The key observation is that $[v]_\kappa$ is highly similar to a symmetric rank-1 tensor, namely

$$[v]_\kappa \approx v^{\otimes \kappa} \in \mathbb{R}^{n \circ \kappa}$$

with the only difference being that $v^{\otimes \kappa}$ contains some terms appearing more than once, which implies that (4.2) could also be casted as an SDP based on the outer product of $v^{\otimes \kappa}$ with itself. Shortly after, we will provide a brief introduction to tensor facts and definitions for clarification. Instead of solving a non-scalable SDP problem for the optimal D over $\mathcal{S}^s$; we propose to apply local search over $\mathbb{R}^{n \circ \kappa}$ for $v^{\otimes \kappa}$, and will analyze when it converges to the global optimum.

For illustration purposes, we first focus on the problem of rank-1 matrix sensing presented in the BM formulation:

$$\min_{x \in \mathbb{R}^n} \quad \|\mathcal{A}(xx^\top - zz^\top)\|_2^2 \tag{4.3}$$

where $M^* = zz^\top$ is the ground truth rank-1 matrix. The objective is to solve (4.3) using a lifted or over-parametrized framework. This means that instead of optimizing over the original vector space $\mathbb{R}^n$, the goal is to optimize over a tensor space, namely $\mathbb{R}^{n \circ l}$ for some $l \geq 2$. Note that (4.3) aims to find a vector x such that

$$\mathcal{A}(xx^\top) = \mathcal{A}(M^*) = \mathcal{A}(zz^\top) := b.$$

Therefore, it is also desirable to achieve $\{\mathcal{A}(xx^\top)\}^{\otimes l} = b^{\otimes l} \in \mathbb{R}^{m \circ l}$. With the repeated application Lemma 4.2.1, we have

$$\{\mathcal{A}(xx^\top)\}^{\otimes l} = \langle \mathbf{A}, x^{\otimes l} \otimes x^{\otimes l} \rangle \tag{4.4}$$

where the tensor $\mathbf{A} \in \mathbb{R}^{(m \circ l) \times (n^2 \circ l)}$ is defined as $A_{m_1 \ldots m_l} = \prod_{k=1}^{l} \otimes \operatorname{vec}(A_{m_k})$. Therefore, one can write the lifted objective similarly to (4.3) as:

$$\min_{\mathbf{w} \in \mathbb{R}^{n \circ l}} \quad \|\langle \mathbf{A}, \mathbf{w} \otimes \mathbf{w} - z^{\otimes l} \otimes z^{\otimes l} \rangle_{m^l+1,\ldots,m^l+n^{2l}}\|_F^2 \tag{4.5}$$

The above derivations can give readers a basic idea of how to construct our lifted tensor objective when $r = 1$. However, when $r > 1$, things become more complicated. A natural extension to general r requires that instead of optimizing over $X \in \mathbb{R}^{n \times r}$, we optimize over $\mathbb{R}^{[n \times r] \circ l}$ tensors, and simply making tensor outer products between $\mathbf{w}$ to be inner products. However, such a tensor space is non-cubical, and subsequently not symmetric. This is the higher-dimensional analogy of non-square matrices, which lacks a number of desirable properties, as per the matrix scenario. In particular, it is necessary for our approach to optimize over a cubical, symmetric tensor space since in the next section we prove that there exists an implicit bias of the gradient descent algorithm under that setting. *Note that in this lifted framework r is the true rank, since the framework itself already offers rich over-parametrization.*

In order to do so, we simply vectorize $X \in \mathbb{R}^{n \times r}$ into $\operatorname{vec}(X) \in \mathbb{R}^{nr}$, and optimize over the tensor space of $\mathbb{R}^{nr \circ l}$, which again is a cubical space. In order to convert a tensor $\mathbf{w} \in \mathbb{R}^{nr \circ l}$ back to $\mathbb{R}^{[n \times r] \circ l}$ to use a meaningful objective, we introduce a new 3-way permutation tensor $\mathbf{P} \in \mathbb{R}^{n \times r \times nr}$ that "unstacks" vectorized matrices. Specifically,

$$\langle \mathbf{P}, \operatorname{vec}(X) \rangle_3 = X \quad \forall X \in \mathbb{R}^{n \times r}, n, r \in \mathbb{Z}^+$$

Such $\mathbf{P}$ can be easily constructed via filling appropriate scalar "1"s in the tensor. Via Lemma 4.2.1, we also know that

$$\langle \mathbf{P}^{\otimes l}, \operatorname{vec}(X)^{\otimes l} \rangle_{3*[l]} = (\langle \mathbf{P}, \operatorname{vec}(X) \rangle_3)^{\otimes l} = X^{\otimes l} \tag{4.6}$$

where $[l]$ denotes the integer set $[1, \dots, l]$, and $c * [l]$ denotes $[c, 2c, \dots, c * l]$ for some $c \in \mathbb{Z}^+$. For notational convenience, we abbreviate $\langle \mathbf{P}^{\otimes l}, \mathbf{w} \rangle_{3*[l]}$ as $\mathbf{P}(\mathbf{w})$ for any arbitrary z-dimensional tensor $\mathbf{w}$ where z can be broken down into the product of two positive integers. Thus, using (4.6), we define our new lifted objective in tensor space:

$$\min_{\mathbf{w} \in \mathbb{R}^{nr \circ l}} \quad \| \langle \mathbf{A}^{\otimes l}, \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{w}) \rangle_{2*[l]} \rangle - b^{\otimes l} \|_F^2 \quad \text{(General lifted formulation)} \tag{4.7}$$

Let us formally define $\mathbf{A}$, $f^l(\cdot) : \mathbb{R}^{n \circ 2l} \mapsto \mathbb{R}$ and $h^l(\cdot) : \mathbb{R}^{[n \times r] \circ l} \mapsto \mathbb{R}$

$$\begin{aligned} &\mathbf{A} \in \mathbb{R}^{m \times n \times n} \quad \text{s.t.} \;\; \mathbf{A}_{kij} = (A_k)_{ij} \; \forall k \in [m], (i,j) \in [n] \times [n] \\ &f^l(\mathbf{M}) := \| \langle \mathbf{A}^{\otimes l}, \mathbf{M} \rangle - b^{\otimes l} \|_F^2 \\ &h^l(\mathbf{w}) = f^l(\langle \mathbf{w}, \mathbf{w} \rangle_{2*[l]}) \end{aligned}$$

with $\nabla f^l(\cdot) = \nabla_{\mathbf{M}} f^l(\cdot)$ and $\nabla h^l(\cdot) = \nabla_{\mathbf{w}} h^l(\cdot)$.

Note that the idea of this lifted formulation (4.7) is that instead of optimizing over a matrix in (1.3), we now optimize over tensors, and the tensor decision variable serves as a surrogate for our original matrix decision variable. When this objective is solved (via whatever optimization algorithm of choice), we perform tensor PCA on this tensor to recover its rank-1 component and extract the underlying low-dimensional matrix/vector. Later in this chapter we show that when using gradient descent, the final tensor will be predominantly rank-1, making this process very easy. Now we offer some basic introduction to tensor definition and basic identities

4.2.1 Tensor Introduction

Definition 7 (Tensor) *As a generalization of the way vectors are used to parametrize finite-dimensional vector spaces, we use* arrays *to parametrize tensors generated from product of finite-dimensional vector spaces, as per [19]. In particular, we define an l-way array as such:*

$$\mathbf{a} = \{ a_{i_1 i_2 \dots i_l} | 1 \leq i_k \leq n_k, 1 \leq k \leq l \} \in \mathbb{R}^{n_1 \times \dots \times n_l}$$

Note that in this paper tensors and arrays can be regarded as synonymous since there exists an isomorphism between them. Moreover, if $n_1 = \dots = n_l$, then we call this tensor(array) an l-order(way), n-dimensional tensor. For the convenience of tensor representation, we use the notation $\mathbb{R}^{n \circ l}$ with $n \circ l := n \times \dots \times n$. In this work, tensors are denoted with bold variables, and other fonts are reserved for matrices, vectors, and scalars unless specified otherwise.

Definition 8 (Symmetric Tensor) *Similar to the definition of symmetric matrices, for an order-l tensor $\mathbf{a}$ with the same dimensions (i.e., $n_1 = \dots = n_l$), also called* *a cubic tensor, it is said that the tensor is symmetric if its entries are invariance under any permutation of their indices:*

$$a_{i_{\sigma(1)}\cdots i_{\sigma(l)}} = a_{i_1\cdots i_l} \quad \forall\sigma, \quad i_1,\dots,i_l \in \{1,\dots,n\}$$

where $\sigma \in \mathcal{G}_l$ denotes a specific permutation and $\mathcal{G}_l$ is the symmetric group of permutations on $\{1,\dots,l\}$. We denote the set of symmetric tensors as $\mathrm{S}^l(\mathbb{R}^n)$.

Definition 9 (Rank of Tensors) *The rank of a cubic tensor $\mathbf{a} \in \mathbb{R}^{n\circ l}$ is defined as*

$$\mathrm{rank}(\mathbf{a}) = \min\{r|\mathbf{a} = \sum_{i=1}^{r} u_i \otimes v_i \otimes \cdots \otimes w_i\}$$

for some vector $u_i,\dots,w_i \in \mathbb{R}^n$. Furthermore, according to [41], if $\mathbf{a}$ is a symmetric tensor, then it can be decomposed as:

$$\mathbf{a} = \sum_{i=1}^{r} \lambda_i u_i \otimes \cdots \otimes u_i := \sum_{i=1}^{r} \lambda_i u_i^{\otimes l}$$

and the rank is conveniently defined as the number of nonzero λ_i's, which is very similar to the rank of symmetric matrices indeed. The most important concept in our paper is rank-1 tensors, and for any tensor $\mathbf{a}$, a necessary and sufficient condition for it to be rank-1 is that

$$\mathbf{a} = u^{\otimes l}$$

for some $u \in \mathbb{R}^n$.

Definition 10 (Tensor Multiplication) *Outer product is an operation carried out on a pair of tensors, denoted as $\otimes$. The outer product of 2 tensors $\mathbf{a}$ and $\mathbf{b}$, respectively of orders l and p, is a tensor of order $l+p$, denoted as $\mathbf{c} = \mathbf{a} \otimes \mathbf{b}$ such that:*

$$c_{i_1\dots i_l j_1\cdots j_p} = a_{i_1\dots i_l} b_{j_1\cdots j_p}$$

When the 2 tensors are of the same dimension, this product is such that $\otimes : \mathbb{R}^{n\circ l} \times \mathbb{R}^{n\circ p} \mapsto \mathbb{R}^{n\circ(l+p)}$. Henceforth, we use the shorthand notation

$$\underbrace{a \otimes \cdots \otimes a}_{l\ times} := a^{\otimes l}$$

We also define an inner product of two tensors. The mode-q inner product between the 2 aforementioned tensors having the same q-th dimension is denoted as $\langle \mathbf{a}, \mathbf{b} \rangle_q$. Without loss of generality, assume that $q = 1$ and

$$\left[\langle \mathbf{a}, \mathbf{b} \rangle_q\right]_{i_2\dots i_l j_2\cdots j_p} = \sum_{\alpha=1}^{n_q} a_{\alpha i_2\dots i_l} b_{\alpha j_2\cdots j_p}$$

Note that when we write $\langle \cdot, \cdot \rangle_q$, we count the q-th dimension of the first entry. Indeed, this definition of inner product can also be trivially extended to multi-mode inner products by just summing over all modes, denoted as $\langle \mathbf{a}, \mathbf{b} \rangle_{q,\dots,s}$.

Lemma 9 (Section 10.2 [67]) *For four arbitrary matrices A, B, C, D of compatible dimensions, it holds that*

$$\langle A \otimes B, C \otimes D \rangle_{2,4} = AC \otimes BD \tag{4.8}$$

4.3 Optimization Landscape of Tensor Formulation

We analyze (4.7) and study the global optimization landscape of this problem. Namely, we will prove that spurious solutions far away from ground truth can be converted into strict saddles with a large enough lifting level l. Note that when $l = 1$, this problem reduces to the original problem (1.3). We start with the characterization of FOPs and SOPs of (4.7). All proofs to this section be can found in Appendix 4.A.

Lemma 10 *The tensor $\hat{\mathbf{w}} \in \mathbb{R}^{nr \circ l}$ is an SOP of* (4.7) *if and only if*

$$\langle \nabla f^l(\langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}}) \rangle_{2*[l]}), \mathbf{P}(\hat{\mathbf{w}}) \rangle_{2*[l]} = 0, \tag{4.9a}$$

$$\begin{aligned} &2\langle \nabla f^l(\langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}}) \rangle_{2*[l]}), \langle \mathbf{P}(\Delta), \mathbf{P}(\Delta) \rangle_{2*[l]} + \\ &\|\langle \mathbf{A}^{\otimes l}, \langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\Delta) \rangle_{2*[l]} + \langle \mathbf{P}(\Delta), \mathbf{P}(\hat{\mathbf{w}}) \rangle_{2*[l]} \rangle\|_F^2 \geq 0 \quad \forall \Delta \in \mathbb{R}^{nr \circ l} \end{aligned} \tag{4.9b}$$

with (4.9b) *being a necessary and sufficient condition for $\hat{\mathbf{w}}$ to be a FOP.*

Now, we turn to showcasing the relationship between the FOPs of (4.7) and those of (1.3), which also have a one-to-one correspondence in the symmetric rank-1 regime. This is the reason why it is necessary to introduce (4.7) despite the extra complication, as rank-1 components tensors in $\mathbb{R}^{[n \times r] \circ l}$ are not lifted versions of $X \in \mathbb{R}^{n \times r}$.

Theorem 3 *For the lifted formulation* (4.7)*, the first-order condition $\nabla h^l(\hat{\mathbf{w}}) = 0$ holds for a symmetric rank-1 tensor $\hat{\mathbf{w}}$ if and only if*

$$\hat{\mathbf{w}} = \text{vec}(\hat{X})^{\otimes l}$$

where $\hat{X} \in \mathbb{R}^{n \times r}$ is an FOP of (1.3)*.*

Theorem 3 establishes a robust connection between the first-order critical points of the lifted formulation and those of the unlifted formulation. This implies that when first-order methods approach a critical point in (4.7), valuable information about an FOP of (1.3) can also be readily extracted. However, the primary challenge in optimizing (1.3) stems from spurious solutions, which cannot be escaped by first or even second-order algorithms. Consequently, it becomes crucial to examine whether the Hessians of the FOPs of (4.7), especially those that correspond to the spurious solutions of (1.3), exhibit any unique properties. As it turns out, the non-global FOPs of (4.7) display some highly favorable characteristics: they no longer constitute second-order critical points of (4.7) and are transformed into strict saddles when the parametrization level l is sufficiently large.

To motivate our analysis of conversion from spurious solutions to strict saddle points, we first offer a closer analysis to the SOPs of the unlifted problem (1.3), which also serves as the key intuition into our main results in this section. The main observation is that, for a spurious SOP $\hat{X}$ and any ground truth Z with $\hat{X}\hat{X}^\top \neq ZZ^\top$, although they all obey conditions (2.5) and (2.6), they still have intrinsic differences that can be amplified via over-parametrization. To illustrate this phenomenon in more detail, we will introduce the following Lemma:

Lemma 11 *For an arbitrary FOP $\hat{X} \in \mathbb{R}^{n\times r}$ of* (1.3) *satisfying the (α_s, r)-RSC property, the following inequality holds:*

$$\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \leq -\alpha_s \frac{\|\hat{X}\hat{X}^\top - M^*\|_F^2}{2\operatorname{tr}(M^*)} \leq 0 \tag{4.10}$$

Now let us recall (2.6), which can be stated equivalently as

$$\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \geq -[\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) \quad \forall U$$

By using the (L_s, r)-RSS property and the assumption that the sensing matrices are symmetric, we can further lower-bound the right-hand side of the above inequality as

$$-[\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) \geq -4[\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top) \geq -4L_s\|\hat{X}U^\top\|_F^2$$

Therefore, it is easy to see that a sufficient condition for the spurious SOPs to disappear is

$$\alpha_s \frac{\|\hat{X}\hat{X}^\top - M^*\|_F^2}{2\operatorname{tr}(M^*)} \geq 4L_s\|\hat{X}U^\top\|_F^2 \quad \forall U \tag{4.11}$$

which means that the L_s and α_s parameters should be benign, and this essentially constitutes the main proof strategy in the existing literature showing in-existence of spurious solutions under benign RIP or RSS/RSC conditions [101, 94, 93, 52, 57].

Therefore, it is natural to ask, in the case when L_s and α_s do not satisfy (4.11), whether one can systematically over-parametrize the problem so that the LHS of (4.11) eventually becomes bigger than the RHS. We know that if we just raise both the RHS and LHS to arbitrary powers, the sign of the inequality will not flip. Therefore, the key insight is that if we keep the constant 4 unchanged, and lift the other terms to arbitrary powers, we can eventually satisfy (2.6). In general terms, we take the following steps in order to establish a strong result regarding the conversion of spurious solutions to strict saddle points:

1. Proving that $\langle \nabla f^l(\langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}})\rangle, \Delta \otimes \Delta\rangle \geq |\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top))|^l$ for some appropriately chosen point $\Delta \in \mathbb{R}^{nr \circ l}$.

2. Proving that $\|\langle \mathbf{A}^{\otimes l}, \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\Delta)\rangle_{2*[l]} + \langle \mathbf{P}(\Delta), \mathbf{P}(\mathbf{w})\rangle_{2*[l]}\rangle\|_F^2 \leq 4L_s\|\hat{X}U^\top\|_F^{2l}$ for some appropriately chosen points $\Delta \in \mathbb{R}^{nr\circ l}$ and $U \in \mathbb{R}^{n\times r}$

3. Finding the smallest l that converts the spurious solution to strict saddle point, under mild technical conditions.

Now we turn to the main result of the general-rank scenario, which concerns the conversion of spurious solutions to strict saddle points. We present the formal results below.

Theorem 4 *Consider an SOP $\hat{X} \in \mathbb{R}^{n\times r}$ of* (1.3) *of general rank $r < n$ with $r_{search} = r$, such that $\hat{X}\hat{X}^\top \neq M^*$, and assume that* (1.3) *satisfies the RSC and RSS conditions. Then $\hat{\mathbf{w}} = \operatorname{vec}(\hat{X})^{\otimes l}$ is a strict saddle of* (4.7) *with a rank-1 symmetric escape direction if $\hat{X}$ satisfies the inequality*

$$\|M^* - \hat{X}\hat{X}^\top\|_F^2 \geq \frac{L_s}{\alpha_s}\lambda_r(\hat{X}\hat{X}^\top)\operatorname{tr}(M^*) \tag{4.12}$$

and l is odd and is large enough so that

$$l > \frac{1}{1-\log_2(2\beta)} \tag{4.13}$$

where β is defined as

$$\beta := \frac{L_s \operatorname{tr}(M^*)\lambda_r(\hat{X}\hat{X}^\top)}{\alpha_s\|M^* - \hat{X}\hat{X}^\top\|_F^2}.$$

Here, L_s and α_s are the respective RSS and RSC constants of (1.3).

Up to this point, we have shown that by lifting the optimization problem (1.3) into tensor spaces, we could convert spurious local solutions into strict saddle points. However, it is also important that we could distinguish the true ground truth solutions $Z \in \mathbb{R}^{n\times r}$ with $ZZ^\top = M^*$ from the spurious ones. This requires that the true solutions Z will remain SOPs after lifting, which we indeed prove in the following theorem:

Theorem 5 *Assume that $Z \in \mathbb{R}^{n\times r}$ is a ground truth solution of* (1.3) *such that $ZZ^\top = M^*$. Then $\operatorname{vec}(Z)^{\otimes l}$ remains an SOP of* (4.7) *regardless of the parametrization level l, and without the need for* (1.3) *to satisfy the RSC or RSS conditions.*

4.4 Implicit Bias in Tensor Optimization

As the readers have observed, the rank-1 constraint on the decision variable $\mathbf{w}$ in Theorem 4 is non-trivial, since finding the dominant rank-1 component of symmetric tensors is itself a non-convex problem in general, and requires a number of assumptions for it to be provably correct [40, 87]. This does not even account for the difficulties of maintaining the symmetric properties of tensors, which also has no natural guarantees. Therefore, although this lifted formulation may be promising in the pursuit of global minimum, there are still major questions to be answered. Most importantly, it is desirable to know *whether the symmetric, rank-1 condition is necessary, and if so, how to achieve it without explicit constraints?*

The necessity of the condition in question can be better understood through insights from [44]. The authors argue that over-parametrizing non-convex optimization problems can reshape the optimization landscape, with the effect being largely independent of the cost function and primarily determined by the parametrization. This notion is consistent with [56], which contends that over-parametrizing vectors into tensors can transform spurious local solutions into strict saddles. However, [44] specifically examines the parametrization from vectors/matrices to tensors, concluding that stationary points are not generally preserved under tensor parametrization, contradicting [56]. This implies that the symmetric, rank-1 constraints required in Theorem 4 are crucial for the conversion of spurious points.

It is essential to devise a method to encourage tensors to be near rank-1, with implicit regularization as a potential solution. There has been a recent surge in examining the implicit regularization effects in first-order optimization methods, such as gradient descent (GD) and stochastic gradient descent (SGD) [45], which has been well-studied in matrix sensing settings [76, 34, 51, 48]. This intriguing observation has prompted us to explore the possible presence of similar implicit regularization in tensor spaces. Our findings indicate that when applying GD to the tensor optimization problem (4.7), an implicit bias can be detected with sufficiently small initialization points. This finding does not directly extend from its matrix counterparts due to the intricate structures of tensors, resulting in a scarcity of useful identities and well-defined concepts for even fundamental properties such as eigenvalues. Furthermore, we show that when initialized at a symmetric tensor, the entire GD trajectory remains symmetric, completing the requirements.

In this section, we study why and how applying gradient descent to (4.7) will result in an implicit bias towards to rank-1 tensors. Prior to presenting the proofs, we shall elucidate the primary intuition behind how GD contributes to the implicit regularization of (1.3). This will aid in comprehending the impact of implicit bias on (4.7), as they share several crucial observations, albeit encountering greater technical hurdles. Consider the first gradient step of (1.3), initialized at a random point $X_0 \in \mathbb{R}^{n \times r_{\text{search}}} = \epsilon X$ with $\|X\|_F^2 = 1$ and $r_{\text{search}} \geq r$:

$$\begin{aligned} X_1 &= X_0 - \eta \nabla h(X_0) = (I + \eta \left[\mathcal{A}^*\mathcal{A}(M^*)\right]) X_0 - \left[\mathcal{A}^*\mathcal{A}(X_0 X_0^\top)\right] X_0 \\ &= (I + \eta \left[\mathcal{A}^*\mathcal{A}(M^*)\right]) X_0 - \epsilon^2 \left[\mathcal{A}^*\mathcal{A}(X X^\top)\right] X_0 \\ &= (I + \eta \left[\mathcal{A}^*\mathcal{A}(M^*)\right]) X_0 + \mathcal{O}(\epsilon^3) \end{aligned}$$

where η is the step-size. Therefore, if ϵ is chosen to be small enough, we have that

$$X_t \approx (I + \eta \mathcal{A}^* \mathcal{A}(M^*))^t X_0 \quad \text{as } \epsilon \to 0$$

Again, according to the symmetric assumptions on $\mathcal{A}$, we can apply spectral theorem on $\mathcal{A}^*\mathcal{A}(M^*) = \sum_{i=1}^n \lambda_i v_i v_i^\top$ for which the eigenvectors are orthogonal to each other. It follows that $X_t \approx \left(\sum_{i=1}^n (1+\eta\lambda_i)^t v_i v_i^\top\right) X_0$.

In many papers surveyed above on making an argument of implicit bias, it is assumed that there is very strong geometric uniformity, or under the context of this paper, it means that $L_s/\alpha_s \approx 1$. Under this assumption, we have $f(M) \approx f(N) + \langle M - N, \nabla f(M)\rangle + \|M-N\|_F^2/2$, leading to the fact that $\nabla^2 f(M) = \mathcal{A}^*\mathcal{A} \approx I$. This immediately gives us $\mathcal{A}^*\mathcal{A}(M^*) \approx M^*$ so that $\lambda_{r+1}, \dots, \lambda_n \approx 0$ as M^* is by assumption a rank-r matrix. This further implies that $X_t \approx \left(\sum_{i=1}^r (1+\eta\lambda_i)^t v_i v_i^\top\right) X_0$, which will become a rank-r matrix, achieving the effect of implicit regularization, as X is now over-parametrized by having $r_{\text{search}} \geq r$.

However, when tackling the implicit regularization problem in tensor space, one key deviation from the aforementioned procedure is that L_s/α_s will be relatively large, as otherwise there will be no spurious solutions, even in the noisy case [101, 53], which is also the motivation for using a lifted framework in the first place. Therefore, instead of saying that $\mathcal{A}^*\mathcal{A}(M^*) \approx M^*$, we aim to show that the gap between the eigenvalues of a comparable tensor term will enlarge as we increase l, making the tensor predominantly rank-1. This observation demonstrates the power of the lifting technique, while at the same time eliminates the critical dependence on a small L_s/α_s factor that is in practice often unachievable due to requiring sample numbers m in the asymptotic regime [14].

Therefore, in order to establish an implicit regularization result for (4.7), there are four major steps that need to be taken:

1. Proving that a point on the GD trajectory $\mathbf{w}_t$ admits a certain breakdown in the form $\mathbf{w}_t = \langle \mathbf{Z}_t, \mathbf{w}_0\rangle - \mathbf{E}_t$ for some $\mathbf{Z}_t$ and $\mathbf{E}_t$.

2. Proving that the spectral norm (equivalence of largest singular value) of $\mathbf{E}_t$ is small (scales with initialization scale ϵ)

3. Proving that $\langle \mathbf{Z}_t, \mathbf{w}_0\rangle$ has a large separation between its largest and second largest eigenvalues using a tensor version of Weyl's inequality.

4. Showing that, with the above holding true, $\mathbf{w}_t$ is predominantly rank-1 after some step t_*.

Lemmas 17, 18, 13, and Theorem 6 correspond to the above four steps, respectively. The reader is referred to Appendix 4.B for the results and for more details.

4.4.1 A Primer on Tensor Algebra and Maintaining Symmetric Property

We start with the spectral norm of tensors, which resembles the operator norm of matrices [69].

Definition 11 *Given a cubic tensor* $\mathbf{w} \in \mathbb{R}^{n^{\circ l}}$*, its spectral norm* $\|\cdot\|_S$ *is defined respectively as:*

$$\|\mathbf{w}\|_S = \sup\left\{|\langle \mathbf{w}, u^{\otimes l}\rangle| \; \|u\|_2 = 1, u \in \mathbb{R}^n\right\}$$

There are many definitions for tensor eigenvalues [68], and in this paper we introduce a novel variational characterization of eigenvalues that resembles the Courant-Fisher minimax definition for eigenvalues of matrices, called the v-Eigenvalue. We denote the i^{th} v-Eigenvalue of $\mathbf{w}$ as $\lambda_i^v(\mathbf{w})$. Note this is a new definition that is first introduced in this paper and might be of independent interest outside of the current scope.

Definition 12 (Variational Eigenvalue of Tensors) *For a given tensor* $\mathbf{w} \in \mathbb{R}^{n^{\circ l}}$*, we define its* k^{th} *variational eigenvalue (v-Eigenvalue)* $\lambda_k^v(\mathbf{w})$ *as*

$$\lambda_k^v(\mathbf{w}) := \max_{\substack{S \\ \dim(S)=k}} \min_{\mathbf{u}\in S} \frac{|\langle \mathbf{w}, \mathbf{u}\rangle|}{\|\mathbf{u}\|_F^2}, \quad k \in [n]$$

where S is a subspace of $\mathbb{R}^{n^{\circ l}}$ that is spanned by a set of orthogonal, symmetric, rank-1 tensors. Its dimension denotes the number of orthogonal tensors that span this space. It is apparent from the definition that $\|\mathbf{w}\|_S = \lambda_1^v(\mathbf{w})$.

Next, since most of our analysis relies on the symmetry of the underlying tensor, it is desirable to show that every tensor along the optimization trajectory of GD on (4.7) remains symmetric if started from a symmetric tensor. Please find its proof in Appendix 4.B.

Lemma 12 *If the GD trajectory of* (4.7) $\{\mathbf{w}_t\}_{t=0}^{\infty}$ *is initialized at a symmetric rank-1 tensor* $\mathbf{w}_0$*, then* $\{\mathbf{w}_t\}_{t=0}^{\infty}$ *will all be symmetric.*

4.4.2 Main Ideas and Proof Sketch

In this subsection, we highlight the main ideas behind implicit bias in GD. Lemma 17 and 18 details the first and second step, and are deferred to Appendix 4.B. The proofs to the results of this section can also be found in that appendix. The lemmas alongside with their proofs are highly technical and not particularly enlightening, therefore omitted here for simplicity. However, the most important takeaway is that for the t^{th} iterate along the GD trajectory of (4.7), we have the decomposition

$$\mathbf{w}_{t+1} = \langle \mathbf{Z}_t, \mathbf{w}_0\rangle - \mathbf{E}_t := \tilde{\mathbf{w}}_t - \mathbf{E}_t$$

for some $\mathbf{Z}_t$ and $\mathbf{E}_t$ such that $\|\mathbf{E}_t\|_S = \mathcal{O}(\epsilon^3)$. This essentially means that by scaling the initialization $\mathbf{w}_0$ to be small in scale, the error term $\mathbf{E}_t$ can be ignored from a spectral standpoint, and scales with ϵ at a cubic rate. This will soon be proven to be useful next.

Lemma 13 *Given $\mathbf{w}_t$ along the GD trajectory of* (4.7)*, its first two v-eigenvalues, as defined in definition 12, satisfy the relation*

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \leq \frac{\|x_0\|_2^l(1+\eta\sigma_2^l(U))^t + \|\mathbf{E}_t\|_S/\epsilon}{|v_1^\top x_0|^l(1+\eta\sigma_1^l(U))^t - \|\mathbf{E}_t\|_S/\epsilon} = \frac{\|x_0\|_2^l(1+\eta\sigma_2^l(U))^t + \mathcal{O}(\epsilon^2)}{|v_1^\top x_0|^l(1+\eta\sigma_1^l(U))^t - \mathcal{O}(\epsilon^2)} \quad (4.14)$$

where $\sigma_1(U)$ and $\sigma_2(U)$ denote the first and second singular values of $U = \langle \mathbf{A}_r^* \mathbf{A}, M^* \rangle \in \mathbb{R}^{nr\times nr}$, and v_1, v_2 are the associated singular vectors.

Lemma 13 showcases that when ϵ is small, the ratio between the largest and second largest v-eigenvalues of $\mathbf{w}$ is dominated by $(\|x_0\|_2^l(1+\eta\sigma_2^l(U))^t)/(|v_1^\top x_0|^l(1+\eta\sigma_1^l(U))^t)$.

Now, if either $\|x_0\|_2^l$ is large or $|v_1^\top x_0|^l$ approaches 0 in value, then the ratio may be relatively large, contradicting our claim. However, this issue can be easily addressed by letting $x_0 = v_1 + g \in \mathbb{R}^{nr}$, where g is a vector with each entry being i.i.d sampled from the Gaussian distribution $\mathcal{N}(0,\rho)$. Note that since $U = \langle \mathbf{A}_r, b\rangle_3$, we can calculate U and v_1 directly. Lemma 19 in Appendix 4.B.2 shows that with this initialization, $|v_1^\top x_0|^l = \mathcal{O}(1)$ and $\|x_0\|_2^l = \mathcal{O}(1)$ with high probability if we select $\rho = \mathcal{O}(1/nr)$. Therefore, the t^{th} iterate along the GD trajectory of (4.7) satisfies

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \asymp \frac{(1+\eta\sigma_2^l(U))^t}{(1+\eta\sigma_1^l(U))^t} \quad (4.15)$$

with hight probability if ρ is small. This implies that "the level of parametrization helps with separation of eigenvalues", since increasing l will decrease ratio $\lambda_2^v(\mathbf{w}_t)/\lambda_1^v(\mathbf{w}_t)$. Furthermore, regardless of the value of $\sigma_1(U)$, a larger t will make this ratio exponentially smaller, proving the efficacy of algorithmic regularization of GD in tensor space.

By combining the above facts, we arrive at a major result showing how a small initialization could make the points along the GD trajectory penalize towards rank-1 as t increases

Theorem 6 *Given the optimization problem* (4.7) *and its GD trajectory over some finite horizon T, i.e., $\{\mathbf{w}_t\}_{t=0}^T$ with $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla h^l(\mathbf{w}_t)$, where η is the stepsize, then there exist $t(\kappa, l) \geq 1$ and $\kappa < 1$ such that*

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \leq \kappa, \qquad \forall t \in [t(\kappa,l), t_T] \quad (4.16)$$

if $\mathbf{w}_0$ *is initialized as* $\mathbf{w}_0 = \epsilon x_0^{\otimes l}$ *with a sufficiently small* ϵ*, where* $t(\kappa, l)$ *is expressed as*

$$t(\kappa, l) = \left\lceil \ln\left(\frac{\|x_0\|_2^l}{\kappa|v_1^\top x_0|^l}\right) \ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1} \right\rceil \tag{4.17}$$

By using the initialization introduced in Lemma 19, we can improve the result of Theoerem 6, which does not need ϵ to be arbitrarily small. The full details are presented in Corollary 1 in Appendix 4.B, stating that as along as

$$t \asymp \ln(1/\kappa) \ln\left((1+\eta\sigma_1^l(U))/(1+\eta\sigma_2^l(U))\right)^{-1}$$

$\mathbf{w}_t$ will be κ-rank-1, as long as ϵ is chosen as a function of U, r, n, L_s, and κ. Note that we say a tensor $\mathbf{w}$ is "κ-rank-1" if $\lambda_2^v(\mathbf{w})/\lambda_1^v(\mathbf{w}) \leq \kappa$.

4.4.3 Approximate Rank-1 Tensors are Benign

Now that we have established the fact that performing gradient descent on (4.7) will penalize the tensor towards rank-1, it begs the question whether approximate rank-1 tensors can also escape from saddle points, which is the most important question under study in this paper. Please find the proofs to the results in this section in Appendix 4.B.

To do so, we first introduce a *major spectral* decomposition of symmetric tensors that is helpful.

Proposition 2 *Given a symmetric tensor* $\mathbf{w} \in \mathbb{R}^{nr \circ l}$*, it can be decomposed into two terms, namely a term consisting of its dominant component and another term that is orthogonal to this direction:*

$$\mathbf{w} = \pm\lambda_1^v(\mathbf{w}) w_s^{\otimes l} + \mathbf{w}^\dagger := \mathbf{w}_\sigma + \mathbf{w}^\dagger, \quad w_s \in \mathbb{R}^n,\ \|w_s\|_2 = 1 \tag{4.18}$$

where $\langle \mathbf{w}, w_s^{\otimes l}\rangle = \lambda_1^v(\mathbf{w})$ *and* $\langle \mathbf{w}^\dagger, w_s^{\otimes l}\rangle = 0$*. Furthermore, if* $\mathbf{w}$ *is a* κ*-rank-1 tensor, then* $\|\mathbf{w}^\dagger\|_S \leq \kappa\lambda_1^v(\mathbf{w}_t)$*.*

Next, we characterize the first-order points of (4.7) with approximate rank-1 tensors in mind. Previously, we showed that if a given FOP of (4.7) is symmetric and rank-1, it has a one-to-one correspondence with FOPs of (1.3). However, if the FOPs of (4.7) are not exactly rank-1, but instead κ-rank-1, it is essential to understand whether they maintain the previous properties. This will be addressed below.

Proposition 3 *Assume that a symmetric tensor* $\mathbf{w} \in \mathbb{R}^{nr \circ l}$ *is an FOP of* (4.7)*, meaning that* (4.9a) *holds. If it is a* κ*-rank-1 tensor with* $\kappa \leq \mathcal{O}(1/\|M^*\|_F^2)$*, then it admits a decomposition as*

$$\mathbf{w} = \pm\lambda_1^v(\mathbf{w})\widehat{w}^{\otimes l} + \mathbf{w}^\dagger$$

with $\mathrm{mat}(\widehat{w}) \in \mathbb{R}^{n\times r}$ *being an FOP of* (1.3) *and* $\|\mathbf{w}^\dagger\|_S \leq \kappa\lambda_1^v(\mathbf{w})$ *by definition.*

The proposition above asserts that for any given FOP of (4.7), if it is κ-rank-1 rather than being truly rank-1, it will consist of a rank-1 term representing a lifted version of an unlifted FOP, as well as a term with a small spectral norm. Referring to (4.55), it is possible to achieve a significantly low κ through a moderate number of iterations. This result, considered the cornerstone of this paper, demonstrates that the use of gradient descent with small initialization will find critical points that are lifted FOPs of (1.3) with added noise, maintaining a robust association between FOPs of (4.7) and (1.3). This finding also facilitates this subsequent theorem:

Theorem 7 *Assume that a symmetric tensor* $\hat{\mathbf{w}} \in \mathbb{R}^{nr \circ l}$ *is an FOP of* (4.7) *that is* κ*-rank-1 with* $\kappa \leq \mathcal{O}(1/\|M^*\|_F^2)$*. Consider its major spectral decomposition* $\hat{\mathbf{w}} = \lambda_S \hat{x}^{\otimes l} + \hat{\mathbf{w}}^\dagger$ *with* $\hat{x} \in \mathbb{R}^{nr}$*, then it has a rank-1 escape direction if* $\hat{X} = \mathrm{mat}(\hat{x})$ *satisfies the inequality*

$$\|M^* - \hat{X}\hat{X}^\top\|_F^2 \geq \frac{L_s}{\alpha_s}\lambda_r(\hat{X}\hat{X}^\top)\,\mathrm{tr}(M^*) + \mathcal{O}(r\kappa^{1/l}) \tag{4.19}$$

where l *is odd and large enough so that* $l > 1/(1-\log_2(2\beta))$ *and* β *is defined as*

$$\beta = \frac{L_s\,\mathrm{tr}(M^*)\lambda_r(\hat{X}\hat{X}^\top)}{\alpha_s\|M^* - \hat{X}\hat{X}^\top\|_F^2 - \mathcal{O}(r\kappa^{1/l})}.$$

This theorem conveys the message that by running GD on (4.7), all critical points have escape directions as long as the point is not close to the ground truth solution.

4.5 Numerical Experiments

4.5.1 A Motivating Example

In this section, we study a class of benchmark matrix sensing instances that have many spurious local minima, where each instance $\mathcal{A}$ is defined as

$$\mathcal{A}_\rho(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij}, & \text{if } (i,j) \in \Omega \\ \rho\mathbf{M}_{ij}, & \text{otherwise} \end{cases}, \tag{4.20}$$

where Ω is a measurement set such that

$$\Omega = \{(i,i),(i,2k),(2k,i) | \quad \forall i \in [n], k \in [\lfloor n/2 \rfloor]\}$$

[88] has proved that each such instance has $\mathcal{O}(2^{\lceil n/2 \rceil} - 2)$ spurious local minima, while it satisfies the RIP property with $\delta_{2r} = (1-\rho)/(1+\rho)$ for some sufficiently small ρ.

To study whether our lifted framework can reshape the optimization landscape of the problem, we analyze the spurious local minima of the unlifted problem (4.3).

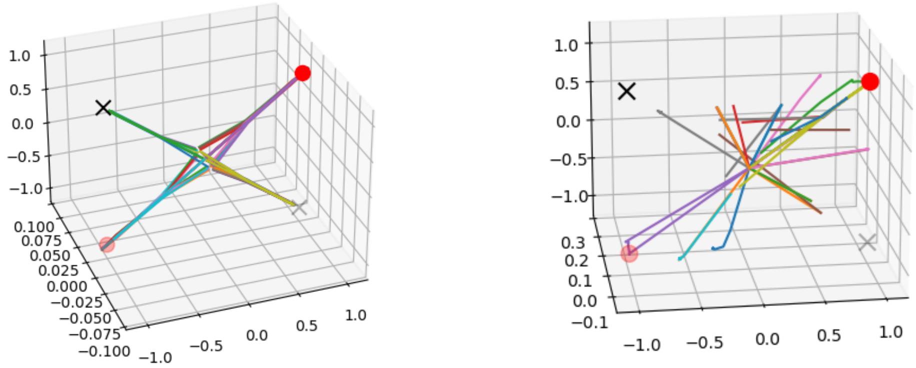
Given any spurious local minimum $\hat{x}$, it is essential to understand whether its lifted counterpart $\hat{x}^{\otimes l}$ behaves differently in (4.5), or more precisely whether $\hat{x}^{\otimes l}$ is still a spurious solution. To get some insight into this question, we conduct numerical experiments to first find the spurious solutions of (4.3) for the measurement matrices given in (4.20), and then find the smallest eigenvalue of the Hessian of (4.5) at the lifted counterpart of each spurious solution. We summarize the findings in Table 4.1 for $\rho = 0.3$. Note that due to the structure of (4.20), the numbers of spurious local minimizers are equal for two cases n and $n+1$ if $\lceil n/2 \rceil = \lceil (n+1)/2 \rceil$, and therefore the results for $n = 4$ and $n = 6$ are omitted.

Table 4.1: The smallest eigenvalue of the Hessian of lifted SOPs of (4.3)

n	l	$\lVert\nabla h^l(z^{\otimes l})\rVert_F$	$\lVert\nabla h^l(\hat{x}^{\otimes l})\rVert_F$	$\lambda_{\min}(\nabla^2 h^l(z^{\otimes l}))$	$\lambda_{\min}(\nabla^2 h^l(\hat{x}^{\otimes l}))$
3	1	0	0	3.99	2.67
3	2	0	0.003	3.99	0.61
3	3	0.004	0.002	3.99	0.24
3	4	0.006	0.004	3.99	-0.17
5	1	0	0	4.18	1.87
5	2	0.002	0	4.56	-0.81
7	1	0.002	0	4.35	1.89
7	2	0.041	0	5.16	-1.64

It can be observed that, for a given spurious local minimizer $\hat{x}$ of (4.3), two properties hold: (i) $\hat{x}^{\otimes l}$ is still a critical point as the gradient of the corresponding objective function h^l is small (its nonzero value is due to the early stopping of the numerical algorithm), (ii) the Hessian at this point becomes smaller as l increases. This means that as the degree of over-parametrization increases, the unlifted spurious local minima will become less of a local minima and more of a strict saddle point. This can be seen for $n = 3$, as every increase in the parametrization leads to a reduced smallest eigenvalue and finally, $\hat{x}^{\otimes l}$ becomes a saddle point with a negative eigenvalue at level $l = 4$, meaning that there is a viable escape direction for gradient descent algorithms. This trend can also be clearly observed for $n = 5$ and $n = 7$, implying that the transformation of the geometric properties at $\hat{x}^{\otimes l}$ is not an isolated phenomenon.

Then we proceed to show the difference in optimization trajectories in both the unlifted and lifted formulations. We choose $\rho = 0.3$ in the numerical experiment, which translates to the RIP constant of $\delta_{2r} = 0.52$, going beyond the known sharp threshold of $\delta < 1/2$, and may create spurious solutions. By the special structure of (4.20), it is easy to verify that there are theoretically 4 SOPs in total, and they



(a) Convergence trajectories of unlifted formulation.

(b) Convergence trajectories of lifted formulation with $l = 3$.

Figure 4.1: The convergence trajectories of (4.20), with $n = 3, \rho = 0.3$. Random gaussian initialization with $\sigma = 0.01, \mu = 0$. 40 Trials in total.

converge to the following 4 points as ρ becomes sufficiently small, which are:

$$\hat{x}_1 \approx \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \hat{x}_2 \approx \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \hat{x}_3 \approx \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \hat{x}_4 \approx \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix}.$$

in which $\hat{x}_1$ and $\hat{x}_4$ are ground truth solutions as $\hat{x}_1\hat{x}_1^\top = \hat{x}_4\hat{x}_4^\top = M^*$. The other SOPs $\hat{x}_2$ and $\hat{x}_3$ are spurious solutions.

To empirically verify that $\rho = 0.3$ is indeed small enough, we simply start from random Gaussian initialization, and apply optimization algorithms to check to what point(s) the algorithm will eventually converge to. We use the standard ADAM optimizer [39] with the hyper-parameter lr $= 0.02$, and the 3D convergence trajectories are plotted in Figure 7.2(a) for 40 different trials with independently sampled initial points. In this plot, the ground truth $\hat{x}_1$ and $\hat{x}_4$ are labeled with big red dots, and $\hat{x}_2$ and $\hat{x}_3$ are labeled with black crosses. One can easily observe that the theoretically derived SOPs are indeed correct, as the plot shows that regardless of initialization, the algorithm will always converge to one of the 4 points given above, which means that $\rho = 0.3$ is already small enough to deteriorate the landscape. Upon a closer scrutiny, one can further realize that all 4 SOPs are equally attractive, and it is impossible to differentiate between ground truth solutions and spurious solutions. In particular, the success rate of applying ADAM to (4.3) with (4.20) is 57.5%. This is highly undesirable in practice because the user will constantly obtain different results by running the same algorithm, leading to confusion as to which result is correct, which exactly represents the inherent difficulty of a highly non-convex optimization problem like (4.20).

Thus, at a high level, it is necessary to show that by using the lifted framework (4.5), we can avoid converging to $\hat{x}_2^{\otimes l}$ and $\hat{x}_3^{\otimes l}$ since with this over-parametrized framework, it is possible that they have become saddle points instead of spurious solutions, as suggested by Theorem 4. To this end, we plot the optimization trajectory of (4.5) with $l = 3$ and (4.20) in Figure 7.2(b), where the optimizer of choice is still ADAM, since it has the ability to escape saddle points and it makes the comparison with Figure 7.2(a) meaningful. The reason that we chose $l = 3$ instead of $l = 2$ is because Theorem 4 only applies to odd values of l. However, one caveat is that since the optimization is performed in tensor space, it is impossible to visualize. To address this issue, instead of showing the full tensor, we perform tensor PCA along each step of the trajectory, and plot the 3D vector that can be transformed to the dominant rank-1 symmetric tensor via tensor outer product. In particular, given a tensor $\mathbf{w}$ on the trajectory, we plot $w \in \mathbb{R}^3$ such that:

$$w = \arg\min_{w} \|\mathbf{w} - w^{\otimes l}\|_F$$

meaning that w is the best projection of $\mathbf{w}$ onto $\mathbb{R}^3$. This is why Figure 7.2(b) seems more complicated than Figure 7.2(a), as an extra layer of approximation is applied. Nevertheless, the message of Figure 7.2(b) is unchanged, as now instead of converging to all 4 points equally, the lifted formulation only converges to the ground truth solutions, as no trajectory leads to the black crosses. This indicates that by converting $\hat{x}_2^{\otimes l}$ and $\hat{x}_3^{\otimes l}$ to saddle points via over-parametrization, we gain real benefits by avoiding spurious solutions, especially compared side-by-side with Figure 7.2(a).

4.5.2 More Experiments

In this subsection, after we run a given algorithm on (4.7) to completion and obtain a final tensor $\mathbf{w}_T$, we then apply tensor PCA (detailed in Appendix 4.C) on $\mathbf{w}_T$ to extract its dominant rank-1 component and recover $X_T \in \mathbb{R}^{n\times r}$ such that $(\mathbf{w}_T)_s = \lambda_s \operatorname{vec}(X_T)^{\otimes l}$. Since $\mathbf{w}_T$ will be approximately rank-1, the success of this operation is expected [40, 87]. We consider a trial to be successful if the recovered X_T satisfies $\|X_T X_T^\top - M^*\|_F \leq 0.05$. We also initialize our algorithm as per Lemma 19.

Perturbed Matrix Completion

As the next step, we apply both lifted and unlifted formulations to (4.20) with $\rho = 0.01$, yielding $\delta_{2r} \approx 1$. We test different values of n and ϵ, using a lifted level of $l = 3$. We ran 10 trials each to calculate success rate. If unspecified in the plot, we default $n = 10$, $\epsilon = 10^{-7}$. Figure 4.2 reveals a higher success rate for the lifted formulation across different problem sizes, with smaller problems performing better as expected (since larger problems require a higher lifting level). Success rates improve with smaller ϵ, emphasizing the importance of small initialization. We employed customGD, a modified gradient descent algorithm with heuristic saddle escaping. This algorithm will deterministically escape from critical points utilizing knowledge from the proof of Theorem 4. For details please refer to Appendix 4.C.

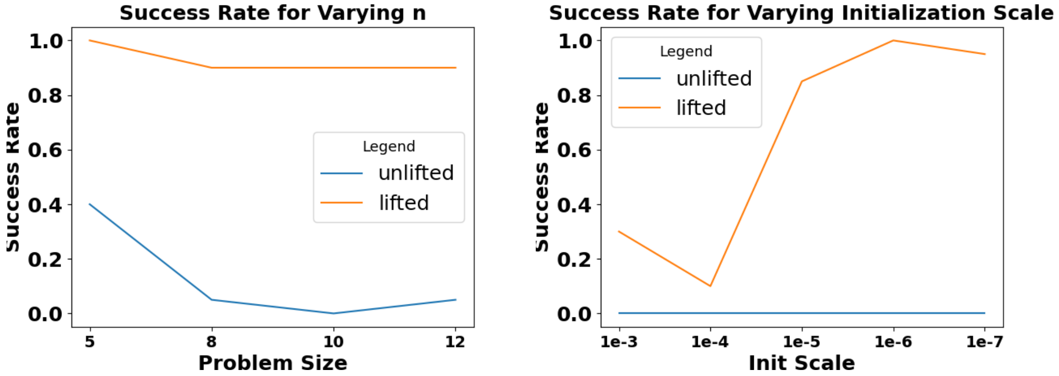


Figure 4.2: Success rate of the lifted formulation versus the unlifted formulation against varying n and ϵ.

Additionally, we examine different algorithms for (4.7), including customGD, vanilla GD, perturbed GD ([33], for its ability to escape saddles), and ADAM [39]. Figure 4.3 suggest that ADAM is an effective optimizer with a high success rate and rapid convergence, indicating that momentum acceleration may not hinder implicit regularization and warrants further research. Perturbed GD performed poorly, possibly due to random noise disrupting rank-1 penalization.

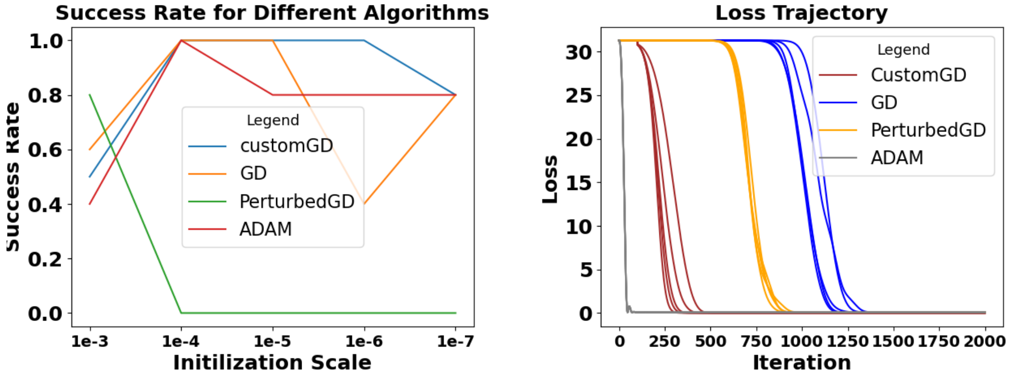


Figure 4.3: Performance of different algorithms applied to the lifted formulation (4.7).

Shallow Neural Network Training with Quadratic Activation

It has long been known that the matrix sensing problem (1.3) includes the training of two-layer neural networks (NN) with quadratic activation as a special case [48].

In summary, the output of the neural network $y \in \mathbb{R}^m$ with respect to m inputs $\{d_i\}_{i=1}^m \in \mathbb{R}^n$ can be expressed as $y_i = \mathbf{1}^\top q(X^\top d_i)$, which implies $y_i = \langle d_i d_i^\top, XX^\top \rangle$, where $q(\cdot)$ is the element-wise quadratic function and $X \in \mathbb{R}^{n\times r}$ in (1.3) represents the weights of the neural network. Thus r represents the number of hidden neurons. In our experiment, we demonstrate that when m is small, the lifted framework (4.7) outperforms standard neural network training in success rate, yielding improved recovery of the true weights. We set the hidden neurons number to be n for the standard network training, thereby comparing the existing over-parametrization framework with the lifted one. We employ the ADAM optimizer for both methods. Table 4.3 showcases the success rate under various problem and sample sizes. Sampling both data and true weights $Z \in \mathbb{R}^{n\times r}$ from an i.i.d Gaussian distribution, we calculate the observations y and attempt to recover Z using both approaches. As the number of samples increases, so does the success rate, with the lifted approach offering significantly better accuracy overall, even when the standard training has a 0% success rate.

Rate	m = 20	m=30	m=40
n=8	0.9(0)	1(0.3)	0.9(0.5)
n=10	0.2(0)	0.6(0)	0.8(0)
n=12	0.1(0)	0.4(0)	0.8(0)

(a) Ground truth weight with $r = 1$

Rate	m = 30	m=40	m=50
n=8	0.3(0)	0.3(0)	0.8(0)
n=10	0.3(0)	0.4(0)	0.2(0)
n=12	0(0)	0(0)	0.2(0)

(b) Ground truth weight with $r = 2$

Table 4.3: Success rate of NN training using (4.7) and original formulation. The number inside the parentheses denotes the success rate of the original formulations. $\epsilon = 10^{-5}$ and $l = 3$.

4.5.3 "Rank-1"ness of Tensors along Optimization Trajectory

In this section, we provide some additional experiments to showcase the algorithmic regularization of GD algorithm in tensor problems like (4.7).

This section involves the decomposition of tensors along the optimization trajectory using a known algorithm, S-HOPM, as outlined in [40]. The S-HOPM algorithms extract the dominant rank-1 component of a given tensor, so as a first step, we apply this to tensors on the trajectory, and obtain $\mathbf{w}_1$. Subsequently, this component was subtracted from the original tensor, and the extraction procedure was repeated on the resultant tensor $\mathbf{w} - \mathbf{w}_1$ to obtain a new component $\mathbf{w}_2$. This allows us to directly compute $\frac{\|\mathbf{w}_1\|_F}{\|\mathbf{w}_2\|_F}$, in the hope to approximate $\lambda_2^v(\mathbf{w}_t)/\lambda_1^v(\mathbf{w}_t)$ for some given t in the trajectory. Note that this procedure mirrors the definition of the variational eigenvalue of tensors defined in Definition 12. The main source of inaccuracy is that the S-HOPM algorithm may not find the real dominant rank-1 component, as specified in the original paper. Therefore, the metric we show below only serves as an approximation of $\lambda_2^v(\mathbf{w}_t)/\lambda_1^v(\mathbf{w}_t)$.

For a practical illustration, we focused on a problem defined in Section 6.1, characterized by a parameter $n = 8$. We were particularly interested in observing the evolution of the aforementioned ratio along the optimization trajectory during the process of gradient descent optimization. The results of this analysis are tabulated below:

iteration	20	40	60	80	100	120	140	160	180
$\epsilon = 10^{-5}$	1.16	0.95	0.82	0.05	0.03	0.018	0.026	0.028	0.013
$\epsilon = 10^{-3}$	0.13	0.43	0.44	0.031	0.036	0.0008	0.034	0.028	0.022
$\epsilon = 0.1$	0.14	0.02	0.05	0.034	0.031	0.026	0.022	0.034	0.037

This table exhibits a notable trend where the tensor gradually exhibits more of a ”rank-1” nature, aligning with the assertions made in Theorem 1. Interestingly, this behavior is observed across varying initialization scales (ϵ), indicating that the phenomenon is not restricted to smaller scales, thus broadening the potential applicability of our findings.

This ratio provides meaningful insights into the training dynamics, which further substantiates the claims made under Theorem 6.

4.6 Summary

This chapter proposed a powerful method to deal with the non-convexity of the matrix sensing problem via the popular BM formulation. Since the problem has several spurious solutions in general and local search methods are prone to be trapped in those points, we developed a new framework via a SOS-type lifting technique to address the issue. We show that although the spurious solutions remain stationary points through the lifting, if a sufficiently rich over-parametrization is used, those spurious solutions will be transformed into strict saddle points (under technical assumptions) and are escapable. This establishes the first result in the literature proving the conversion of spurious solutions to saddle points, and it quantifies how much over-parametrization is needed to break down the complexity of the problem.

Our study also highlights the pivotal role of gradient descent in inducing implicit regularization within tensor optimization, specifically in the context of the lifted matrix sensing framework. We reveal that GD can lead to approximate rank-1 tensors and critical points with escape directions when initialized at an adequately small scale. This work also contributes to the usage of tensors in machine learning models, as we introduce novel concepts and techniques to cope with the intrinsic complexities of tensors.

4.A Missing Details of Section 4.3

Proof 7 (Proof of Lemma 10) *We have*

$$\nabla f^l(\mathbf{M}) = \langle\langle \mathbf{A}^{\otimes l}, \mathbf{M} - \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})\rangle, \mathbf{A}^{\otimes l}\rangle_{1,4,\dots,3l-2} \tag{4.21}$$

where the new map $\mathcal{M} : \mathbb{R}^{nr\circ l} \mapsto \mathbb{R}^{n\circ 2l}$ *is defined as*

$$\mathcal{M}(\mathbf{w}) = \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{w})\rangle_{2*[l]},$$

and its total derivative at $\mathbf{w}$ *is the linear map* $D_{\mathbf{w}}\mathcal{M} : \mathbb{R}^{nr\circ l} \mapsto \mathbb{R}^{n\circ 2l}$ *given below:*

$$D_{\mathbf{w}}\mathcal{M}(\mathbf{v}) = \langle \mathbf{P}(\mathbf{v}), \mathbf{P}(\mathbf{w})\rangle_{2*[l]} + \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{v})\rangle_{2*[l]}. \tag{4.22}$$

Combining (4.21) *and* (4.22) *gives that*

$$D_{\mathbf{w}}h^l(\mathbf{v}) = \langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}}\mathcal{M}(\mathbf{v})\rangle^\top \langle \mathbf{A}^{\otimes l}, \mathcal{M}(\mathbf{w}) - \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})\rangle \tag{4.23}$$

The sensing matrices A_k $\forall k \in [m]$ *are assumed to be symmetric, and therefore* $\langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}}\mathcal{M}(\mathbf{v})\rangle = 2\langle \mathbf{A}^{\otimes l}, \langle \mathbf{P}(\mathbf{v}), \mathbf{P}(\mathbf{w})\rangle_{2*[l]}\rangle$.

Therefore, since the first-order optimality condition for (4.7) *is that* $D_{\mathbf{w}}h^l(\mathbf{v}) = 0$ $\forall \mathbf{v} \in \mathbb{R}^{nr\circ l}$*, it can be equivalently written as*

$$\langle\langle \mathbf{A}^{\otimes l}, \mathbf{P}(\mathbf{w})\rangle_{2*[l]}, \langle \mathbf{A}^{\otimes l}, \mathcal{M}(\mathbf{w}) - \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})\rangle\rangle_{1,3,\dots,2l-1} = 0, \tag{4.24}$$

and left-hand side of the above equation yields (4.9a) *after rearrangements.*

For the second-order optimality condition, one can directly take the derivative of $D_{\mathbf{w}}h^l(\mathbf{v})$*, but there is an easier way since we are only concerned the expression of its quadratic form evaluated at some tensor* $\Delta \in \mathbb{R}^{nr\circ l}$*. For a brief moment, assume that we aim to optimize over* $\mathbf{X} \in \mathbb{R}^{[n\times r]\circ l}$*, for which*

$$\nabla h^l(\mathbf{X}) = 2\langle \nabla f^l(\langle \mathbf{X}, \mathbf{X}\rangle_{2*[l]}), \mathbf{X}\rangle_{2*[l]} \in \mathbb{R}^{[n\times r]\circ l}$$

Therefore, if we instead take the derivate of $g(\mathbf{P}(\mathbf{w}))$ *with respect to* $\mathbf{w}$*, we can simply use the chain rule and arrive at*

$$\nabla_{\mathbf{w}} h^l(\mathbf{P}(\mathbf{w})) = \langle \nabla h^l(\mathbf{X}), \mathbf{P}^{\otimes l}\rangle_{1,2,4,5,\dots,3l-1,3l} \tag{4.25}$$

Hence, if we take the derivate of ∇h^l *and evaluate it at* $\mathbf{X}$ *in the direction of* $\mathbf{U} \in \mathbb{R}^{[n\times r]\circ l}$*, we obtain that*

$$\begin{aligned} D_{\mathbf{X}}\nabla h^l(\mathbf{U}) = {} & 2\langle \nabla f^l(\langle \mathbf{X}, \mathbf{X}\rangle_{2*[l]}), \mathbf{U}\rangle_{2*[l]} + \langle\langle \mathbf{A}^{\otimes l}, \langle \mathbf{X}, \mathbf{U}\rangle_{2*[l]} + \langle \mathbf{U}, \mathbf{X}\rangle_{2*[l]}\rangle, \langle \mathbf{A}^{\otimes l}, \mathbf{w}\rangle_{2,5,\dots,3l-1}\rangle \\ & + \langle\langle \mathbf{A}^{\otimes l}, \langle \mathbf{X}, \mathbf{U}\rangle_{2*[l]} + \langle \mathbf{U}, \mathbf{X}\rangle_{2*[l]}\rangle, \langle \mathbf{A}^{\otimes l}, \mathbf{w}\rangle_{3,6,\dots,3l}\rangle \end{aligned}$$

Combined with (4.25)*, we conclude that*

$$\begin{aligned} [\nabla^2_{\mathbf{w}} h^l(\mathbf{P}(\mathbf{w}))](\mathbf{v}, \mathbf{v}) &= 2\langle \nabla f^l(\mathcal{M}(\mathbf{w})), \mathcal{M}(\mathbf{v})\rangle + \langle\langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}}\mathcal{M}(\mathbf{v})\rangle, \langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}}\mathcal{M}(\mathbf{v})\rangle\rangle \\ &= 2\langle \nabla f^l(\mathcal{M}(\mathbf{w})), \mathcal{M}(\mathbf{v})\rangle + \|\langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}}\mathcal{M}(\mathbf{v})\rangle\|_F^2 \end{aligned}$$

which yields (4.9b) *directly.*

Proof 8 (Proof of Theorem 3) *When* $\hat{\mathbf{w}} = \mathrm{vec}(\hat{X})^{\otimes l}$, *Lemma 4.2.1 and* (4.9a) *together imply that*

$$\langle \nabla f^l(\langle \hat{X}^{\otimes l}, \hat{X}^{\otimes l}\rangle_{2*[l]}), \hat{X}^{\otimes l}\rangle_{2*[l]} = (\nabla f(\hat{X}\hat{X}^\top)\hat{X})^{\otimes l} = 0 \tag{4.26}$$

which is equivalent to

$$\nabla f(\hat{X}\hat{X}^\top)\hat{X} = 0,$$

which is exactly (2.5).

Proof 9 (Proof for Lemma 11) *According to [94],* $\nabla f(M)$ *can be assumed to be symmetric without loss of generality. Hence, one can select* $u \in \mathbb{R}^n$ *such that* $u^\top \nabla f(\hat{x}\hat{x}^\top)u = \lambda_{min}(\nabla f(\hat{x}\hat{x}^\top))$. *Then via the definition of RSC we have*

$$f(M^*) \geq f(\hat{X}\hat{X}^\top) + \langle \nabla f(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top\rangle + \frac{\alpha_s}{2}\|\hat{X}\hat{X}^\top - M^*\|_F^2.$$

Given that $\hat{X}$ *is also an FOP, we have that*

$$\langle \nabla f(\hat{X}\hat{X}^\top), \hat{X}\hat{X}^\top\rangle = 0$$

according to (2.5) *and since* $f(\hat{X}\hat{X}^\top) - f(M^*) \geq 0$, *one can write that*

$$\langle \nabla f(\hat{X}\hat{X}^\top), M^*\rangle \leq -\frac{\alpha_s}{2}\|\hat{x}\hat{x}^\top - M^*\|_F^2$$

after rearrangements. Furthermore, since both $\nabla f(\hat{X}\hat{X}^\top)$ *and* M^* *are assumed to be positive semidefinite for the above-mentioned reasons, we have that*

$$\langle \nabla f(\hat{X}\hat{X}^\top), M^*\rangle \geq \lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top))\,\mathrm{tr}(M^*)$$

which implies that

$$\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \leq -\alpha_s \frac{\|\hat{X}\hat{X}^\top - M^*\|_F^2}{2\,\mathrm{tr}(M^*)} \leq 0 \tag{4.27}$$

This completes the proof.

Proof 10 (Proof of Theorem 4) *By Lemma 11, we select* $u \in \mathbb{R}^n$ *such that*

$$u^\top \nabla f(\hat{X}\hat{X}^\top)u = \lambda_{min}(\nabla f(\hat{X}\hat{X}^\top))$$

with $\lambda_{min}(\nabla f(\hat{X}\hat{X}^\top)) \leq 0$. *Now define* $G \coloneqq -\lambda_{min}(\nabla f(\hat{X}\hat{X}^\top)) \geq 0$. *If we label*

$$C_1 \coloneqq \langle \nabla f(\hat{X}\hat{X}^\top), UU^\top\rangle, \quad C_2 \coloneqq [\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top, \hat{X}U^\top)$$

Then we have that $C_1 = -G$. *Also, since the sensing matrices* A_a *can be assumed be to symmetric, we have that*

$$[\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) = 4[\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top, \hat{X}U^\top).$$

Additionally we choose $q \in \mathbb{R}^r$ *to be the* r*-th singular value of* $\hat{X}$, *with*

$$\|\hat{X}q\|_2 = \sigma_r(\hat{X}), \qquad \|q\|_2 = 1$$

and define $U \in \mathbb{R}^{n\times r} = uq^\top$. *Subsequently, the RSS condition can be used to show that*

$$\begin{aligned}&[\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) \leq L_s\|\hat{X}U^\top + U\hat{X}^\top\|_F^2\\ &= L_s\|u(\hat{X}q)^\top + (\hat{X}q)u^\top\|_F^2 = 2L_s\|\hat{X}q\|_F^2 + 2L_s(q^\top(\hat{X}^\top u))^2 = 2L_s\lambda_r(\hat{X}\hat{X}^\top)\end{aligned}$$

since $\hat{X}^\top u = 0$ *according to the first-order condition* (2.5). *Therefore,*

$$C_2 \leq \frac{1}{2}L_s\lambda_r(\hat{X}\hat{X}^\top)$$

Now, if we choose $\Delta = \mathrm{vec}(U)^{\otimes l}$ *for the aforementioned* $U \in \mathbb{R}^{n\times r}$, *the LHS of* (4.9b) *can be expressed as:*

$$\begin{aligned}LHS &= 2(\langle \mathbf{A}, \hat{X}\hat{X}^\top\rangle_{2,3}^\top\langle \mathbf{A}, uu^\top\rangle_{2,3})^l - 2(\langle \mathbf{A}, M^*\rangle_{2,3}^\top\langle \mathbf{A}, uu^\top\rangle_{2,3})^l + 4(\|\langle \mathbf{A}, \hat{X}U^\top\rangle_{2,3}\|_2^2)^l\\ &\leq 2(\lambda_{min}(\nabla f(\hat{X}\hat{X}^\top)))^l + 4C_2^l\\ &= 2C_1^l + 4C_2^l\end{aligned} \tag{4.28}$$

where the inequality follows from:

$$a^n - b^n \leq (a-b)^n, \quad \forall b \geq a \geq 0$$

Here, since $a - b = C_1 \leq 0$, *the above inequality can be used. As a result,*

$$LHS\ of\ (4.9\mathrm{b}) \leq \underbrace{-2G^l}_{Part\ 1} + \underbrace{\frac{2}{2^{l-1}}L_s^l\lambda_r(\hat{X}\hat{X}^\top)^l}_{Part\ 2}$$

We know since $G \geq 0$, *Part 1 is always negative assuming* l *is odd, and Part 2 is always positive. Therefore, it suffices to find an order* l *such that*

$$G^l > (1/2^{l-1})L_s^l\lambda_r(\hat{X}\hat{X}^\top)^l \tag{4.29}$$

To derive a sufficient condition for (4.29), *we first need a lower bound on* G, *and Lemma* (11) *conveniently provides this bound, giving that*

$$G \geq \frac{\alpha_s}{2\,\mathrm{tr}(M^*)}\|M^* - \hat{X}\hat{X}^\top\|_F^2 \tag{4.30}$$

Therefore, if

$$\left(\frac{\alpha_s}{2\,\mathrm{tr}(M^*)}\|M^* - \hat{X}\hat{X}^\top\|_F^2\right)^l > (1/2^{l-1})L_s^l \lambda_r(\hat{X}\hat{X}^\top)^l,$$

we can conclude that (4.29) *holds, which implies that the LHS of* (4.9b) *is negative, directly proving that* $\hat{X}^{\otimes l}$ *is not an SOP anymore. Elementary manipulations of the above equation give that a sufficient condition is*

$$\|M^* - \hat{X}\hat{X}^\top\|_F^2 > 2^{1/l}\frac{L_s}{\alpha_s}\lambda_r(\hat{X}\hat{X}^\top)\,\mathrm{tr}(M^*) \tag{4.31}$$

We now consider (4.12)*, which means that*

$$\lambda_r(\hat{X}\hat{X}^\top) \leq \frac{\alpha_s}{L_s\,\mathrm{tr}(M^*)}\|M^* - \hat{X}\hat{X}^\top\|_F^2 \tag{4.32}$$

Subsequently, define a constant γ *such that*

$$L_s\lambda_r(\hat{X}\hat{X}^\top) = \gamma(\frac{\alpha_s}{2\,\mathrm{tr}(M^*)}\|M^* - \hat{X}\hat{X}^\top\|_F^2)$$

Then, according to Lemma 1 and (4.30)*, we can conclude that* $\gamma \geq 1$*. Moreover,* (4.32) *also means that* $\gamma < 2$*. With this new definition, the sufficient condition* (4.31) *becomes*

$$1 > \frac{\gamma}{2^{(l-1)/l}} \tag{4.33}$$

Since we already know that $1 \leq \gamma < 2$*, there always exists a large enough* l *such that* (4.33) *holds, which in turn implies that LHS of* (4.9b) *is negative, proving that* $\mathrm{vec}(\hat{X})^{\otimes l}$ *is a saddle point with the escape direction* $\mathrm{vec}(U)^{\otimes l}$*, proving the claim.*

Next, we aim to study how large l *needs to be in order for* (4.33) *to hold. Again, we know that*

$$\gamma = \frac{2L_s\,\mathrm{tr}(M^*)\lambda_r(\hat{X}\hat{X}^\top)}{\alpha_s\|M^* - \hat{X}\hat{X}^\top\|_F^2} := 2\beta$$

and that $\beta \leq 1$ *due to assumption* (4.12)*. Therefore, for* (4.33) *to hold true, it is enough to have*

$$2^{(l-1)/l} > 2\beta \implies \frac{l-1}{l} > \log_2(2\beta) \implies l > \frac{1}{1-\log_2(2\beta)}$$

Proof 11 (Proof of Theorem 5) *Let us start with the first-order optimality condition. Consider the linear map in the proof of Lemma 10* $\mathcal{M} : \mathbb{R}^{nr\circ l} \mapsto \mathbb{R}^{n\circ 2l}$

$$\mathcal{M}(\mathbf{w}) = \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{w})\rangle_{2*[l]},$$

Again, it is apparent that

$$\nabla f^l(\mathbf{M}) = \langle\langle \mathbf{A}^{\otimes l}, \mathbf{M} - \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})\rangle, \mathbf{A}^{\otimes l}\rangle_{1,4,\ldots,3l-2}$$

Therefore, at the point $\mathbf{M} = \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})$, *we know that* $\nabla f^l(\mathcal{M}(\mathrm{vec}(Z)^{\otimes l})) = 0$. *Consequently, the LHS of* (4.9a) *is equal to zero since it is a product between* $\nabla f^l(\mathcal{M}(\mathrm{vec}(Z)^{\otimes l}))$ *and* $\mathbf{P}(\mathrm{vec}(Z)^{\otimes l})$.

Next, we turn to the second-order optimality condition. Again, recall from the proof of Lemma 10 that

$$\textit{LHS of } (4.9\mathrm{b}) = \underbrace{2\langle \nabla f^l(\mathcal{M}(\mathbf{w})), \mathcal{M}(\Delta)\rangle}_{\textit{Part 1}} + \underbrace{\|\langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}}\mathcal{M}(\Delta)\rangle\|_F^2}_{\textit{Part 2}}$$

By the above arguments, we have $\nabla f^l(\mathcal{M}(\mathbf{w})) = 0$ *when* $\mathbf{w} = \mathrm{vec}(Z)^{\otimes l}$, *meaning that Part 1 equals to zero. This implies that*

$$\textit{LHS of } (4.9\mathrm{b}) = \|\langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}}\mathcal{M}(\Delta)\rangle\|_F^2 \geq 0, \qquad \forall \Delta$$

regardless of the values of $\mathbf{A}$ *or* $\mathbf{w} = \mathrm{vec}(Z)^{\otimes l}$.

4.B Missing Details of Section 4.4

4.B.1 More Tensor Algebra

Definition 13 *Given a cubic tensor* $\mathbf{w} \in \mathbb{R}^{n \circ l}$, *its spectral norm* $\|\cdot\|_S$ *and nuclear norm* $\|\cdot\|_*$ *are defined respectively as*

$$\|\mathbf{w}\|_* = \inf\left\{\sum_{j=1}^{r_m} |\lambda_j| : \mathbf{w} = \sum_{j=1}^{r_m} \lambda_j w_j^{\otimes l},\ \|w_j\|_2 = 1, w_j \in \mathbb{R}^n\right\}$$
$$\|\mathbf{w}\|_S = \sup\left\{|\langle \mathbf{w}, u^{\otimes l}\rangle|\ \|u\|_2 = 1, u \in \mathbb{R}^n\right\}$$

From the definition, it also follows that

$$\|\mathbf{w}\|_S \leq \|\mathbf{w}\|_*$$

The above definitions are similar to those for their matrix counterparts. However, unlike the spectral norm of matrices, the spectral norm of tensors are not tensor norms, namely that they do not obey

$$\|\langle \mathbf{w}, \mathbf{v}\rangle\|_S \leq \|\mathbf{w}\|_S \|\mathbf{v}\|_S$$

in general. Conversely, the nuclear norm is a valid tensor norm, and we have the following property:

Lemma 14 (Theorem 2.1, 3.2 [69]) *For tensors* $\mathbf{w}$ *and* $\mathbf{v}$ *of appropriate dimensions (if doing inner product, the dimensions along which the multiplication is performed must have matching size), we have*

$$\|\langle \mathbf{w}, \mathbf{v} \rangle\|_S \leq \|\mathbf{w}\|_S \|\mathbf{v}\|_*$$
$$\|\langle \mathbf{w}, \mathbf{v} \rangle\|_* \leq \|\mathbf{w}\|_* \|\mathbf{v}\|_*$$

Moreover, they have a dual norm relationship:

Lemma 15 (Lemma 21 [49]) *The spectral norm* $\|\cdot\|_S$ *is the dual norm to the nuclear norm* $\|\cdot\|_*$*, namely given an arbitrary tensor* $\mathbf{w}$*, we have that*

$$\|\mathbf{w}\|_S = \sup_{\|\mathbf{v}\|_* \leq 1} |\langle \mathbf{w}, \mathbf{v} \rangle|$$

with $\mathbf{v}$ *having the same dimensions as* $\mathbf{w}$*.*

Next, we introduce the notion of eigenvalues for tensors. There are many related definitions, like outlined in [68]. However, we introduce a novel variational characterization of eigenvalues that resembles the Courant-Fisher minimax definition for eigenvalues of matrices, stated in Definition 12. Note this is a new definition that is first introduced in this paper, and may be of independent interest outside of the current scope.

It is apparent from the definition that $\|\mathbf{w}\|_S = \lambda_1^v(\mathbf{w})$. Note that our definition of v-Eigenvalues of tensors can only define n eigenvalues at most, which is not the maximum amount of H- or Z-Eigenvalues a tensor can have [68], and it is well known that even with symmetric tensors, its rank can go well beyond n [19]. We also note that this definition exactly coincides with the definition of Hermitian tensor eigenvalues (introduced here [63]) when constrained to Hermitian tensors [17]. We also conjecture that this definition coincides with the top-n Z-Eigenvalues for even-order symmetric real tensors [68], but it is an open question for now.

Using the definition of v-Eigenvalues, we can also obtain an equivalent characterization, just like the Courant-Fisher definition for matrix eigenvalues, which helps us in proving a tensor version of Weyl's inequality:

Proposition 4 *For an integer* k *in* $[1, \dots, n]$*, the* k^{th} *variational eigenvalue (v-Eigenvalue)* $\lambda_k^v(\mathbf{w})$ *of a tensor* $\mathbf{w}$ *satisfies:*

$$\lambda_k^v(\mathbf{w}) = \min_{\substack{T \\ \dim(T)=n-k+1}} \max_{\mathbf{u} \in T} \frac{|\langle \mathbf{w}, \mathbf{u} \rangle|}{\|\mathbf{u}\|_F^2} = \max_{\substack{S \\ \dim(S)=k}} \min_{\mathbf{u} \in S} \frac{|\langle \mathbf{w}, \mathbf{u} \rangle|}{\|\mathbf{u}\|_F^2}$$

Proof 12 (Proof of Proposition 4) *We prove the proposition by contradiction. Assume that the two formulations claimed to be identical in Proposition 4 are not the same. We further assume that* S *is spanned by symmetric, rank-1 tensors* $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$*,* *and that T is spanned by symmetric, rank-1 tensors $\{\mathbf{u}_{-(n-k+1)}, \dots, \mathbf{u}_{-1}\}$, meaning that*

$$\langle \mathbf{w}, \mathbf{u}_k \rangle \neq \langle \mathbf{w}, \mathbf{u}_{-(n-k+1)} \rangle$$

assuming that $\mathbf{u}_k$ and $\mathbf{u}_{-(n-k+1)}$ are the inner *argmin and argmax of their respective formulations with norm 1. Since they have to be rank-1 tensors (if not we can decrease the proportion of orthogonal elements with higher or lower $|\langle \mathbf{w}, \mathbf{u} \rangle|$ values), it is possible to denote*

$$\mathbf{u}_k = u_k^{\otimes l}, \quad \mathbf{u}_{-(n-k+1)} = u_{-(n-k+1)}^{\otimes l} \textit{ where } u_k, u_{-(n-k+1)} \in \mathbb{R}^n$$

We also know that u_k and $u_{-(n-k+1)}$ are linearly independent, as otherwise $\mathbf{u}_k$ and $\mathbf{u}_{-(n-k+1)}$ will have the same inner product with $\mathbf{w}$. Thus, assume

$$u_k = \xi_1 u_{-(n-k+1)} + \xi_2 u_{-(n-k+1)}^{\perp}, \quad \xi_2 \neq 0.$$

It follows that

$$\mathbf{u}_k = \xi_1^l u_{-(n-k+1)}^{\otimes l} + \xi_2^l (u_{-(n-k+1)}^{\perp})^{\otimes l} + \underbrace{\dots\dots}_{\textit{other non-symmetric terms}}$$

Denote $(u_{-(n-k+1)}^{\perp})^{\otimes l} := \mathbf{u}_{k+1}$. Now, it follows from definition that

$$\mathbf{u}_{k+1} \perp \{\mathbf{u}_1, \dots, \mathbf{u}_{k-1}\}$$

and also

$$\mathbf{u}_{k+1} \notin \textit{span}\{\mathbf{u}_{-(n-k)}, \dots, \mathbf{u}_{-1}\}$$

as otherwise the outer *maximization formulation affecting the choice of u_k will make $\xi_2 = 0$, contradicting our claim. By definition we have*

$$\textit{span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\} \bigcap \textit{span}\{\mathbf{u}_{-(n-k)}, \dots, \mathbf{u}_{-1}\} = \{\emptyset\}$$

In summary we have that $\mathbf{u}_{k+1} \perp \mathbf{u}_{-(n-k+1)}, \{\mathbf{u}_1, \dots, \mathbf{u}_{k-1}\}, \{\mathbf{u}_{-(n-k)}, \dots, \mathbf{u}_{-1}\}$, meaning that we have obtained $n+1$ symmetric rank-1 and n-dimensional tensors all orthogonal to each other, which is apparently not possible, thus refuting our initial claim.

With this new definition equipped, we proceed to show a tensor version of Weyl's inequality, which is key in our proof as promised.

Lemma 16 (Tensor Weyl's) *Consider two tensors $\mathbf{w}$ and $\mathbf{v}$ of the same dimension. It holds that*

$$\lambda_k^v(\mathbf{w}) + \lambda_1^v(\mathbf{v}) \geq \lambda_k^v(\mathbf{w} + \mathbf{v}) \geq \lambda_k^v(\mathbf{w}) - \lambda_1^v(\mathbf{v}) \tag{4.34}$$

The proof of Lemma 16 is highly similar to that of Theorem 2 in [17], only substituting for our new definition of v-Eigenvalues, thus omitted for simplicity.

4.B.2 Main Results and Their Proofs

Note that in this section some tensor inner products will be written as if they were matrices for clarity of writing, and some subscripts for inner-products will be dropped when obvious. If two tensors in $\mathbb{R}^{nr\circ 2l}$ are multiplied together, then the even dimensions of the first tensor will be inner-producted with the odd dimensions of the second tensor. When a tensor in $\mathbb{R}^{nr\circ 2l}$ multiplies with a tensor in $\mathbb{R}^{nr\circ l}$, then the even dimensions of the first tensor will be inner-producted with all the dimensions of the second tensor.

We start with the proof to Lemma 12.

Proof 13 (Proof of Lemma 12) *We proceed with the proof by induction. First, assume that $\mathbf{w}_0 = x_0^{\otimes l}$ for some $x_0 \in \mathbb{R}^{nr}$. One can write*

$$\nabla h^l(\mathbf{w}_0) = \langle\langle (I_r \oslash_{1,2} \mathbf{A})^{\otimes l}, \mathbf{w}_0\rangle_{2*[l]}, \langle \mathbf{A}^{\otimes l}, \mathcal{M}(\mathbf{w}_0) - \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})\rangle\rangle_{1,3,\dots,2l-1} \tag{4.35}$$

where $\mathcal{M}(\cdot)$ is defined per proof of Lemma 10. The difference between this formulation and (4.24) *is that we have replaced $\langle \mathbf{A}^{\otimes l}, \mathbf{P}(\mathbf{w}_0)\rangle_{2*[l]}$ with $\langle (I_r \oslash_{1,2} \mathbf{A})^{\otimes l}, \mathbf{w}_0\rangle_{2*[l]}$, which are equivalent, just with the second tensor having the dimensions $nr, m, \dots, nr, m$ so that $\nabla h^l(\mathbf{w}_0)$ has the dimensions $nr, \dots, nr$. Note that $\oslash$ denotes the usual kronecker product, which can be thought of a reshaped version of tensor outer product. $\oslash_{1,2}$ denotes the kronecker product only happening with respect to the first 2 dimensions of $\mathbf{A}$. From now on, we denote $\mathbf{A}_r := I_r \oslash_{1,2} \mathbf{A}$.*

Now, according to the above formulation and Lemma 4.2.1, we have

$$\begin{aligned}\nabla h^l(\mathbf{w}_0) &= \left(\langle \mathbf{A}_r, \langle \mathbf{A}, \mathrm{mat}(x_0)\,\mathrm{mat}(x_0)^\top - M^*\rangle\rangle_{3,6,\dots,3l}\ x_0\right)^{\otimes l} \\ &:= \left(\langle \mathbf{A}_r^*\mathbf{A}, \mathrm{mat}(x_0)\,\mathrm{mat}(x_0)^\top - M^*\rangle\ x_0\right)^{\otimes l}\end{aligned} \tag{4.36}$$

where

$$(\mathbf{A}_r^l)^*\mathbf{A}^l := \langle (\mathbf{A}_r)^{\otimes l}, \mathbf{A}^{\otimes l}\rangle_{3,6,\dots,3l} \in \mathbb{R}^{[nr\times nr\times n\times n]\circ l} \tag{4.37}$$

Now, $\langle \mathbf{A}_r^*\mathbf{A}, \mathrm{mat}(x_0)\,\mathrm{mat}(x_0)^\top - M^*\rangle$ is an $nr \times nr$ matrix, so the above tensor is simply a vector outer product, being symmetric by definition. Consequently, $\mathbf{w}_1 = \mathbf{w}_0 - \eta\nabla h^l(\mathbf{w}_0)$ is still symmetric, since the addition of symmetric tensors maintains symmetric property. This completes the proof of the initial step.

Then, we proceed to show the induction step. Assume that $\mathbf{w}_{t-1}$ is symmetric, meaning that

$$\mathbf{w}_{t-1} = \sum_{j=1}^{r_m} \lambda_j (x_j^{t-1})^{\otimes l}, \quad x_j^{t-1} \in \mathbb{R}^{nr}$$

where r_m is the symmetric rank of $\mathbf{w}_{t-1}$. This means that

$$\nabla h^l(\mathbf{w}_{t-1}) = \sum_{j_1,j_2,j_3}^{r_m,r_m,r_m} \lambda_{j_1}\lambda_{j_2}\lambda_{j_3}(\langle \mathbf{A}_r^*\mathbf{A}, \operatorname{mat}(x_{j_1}^{t-1})\operatorname{mat}(x_{j_2}^{t-1})^\top\rangle x_{j_3}^{t-1})^{\otimes l} - \\ \sum_{j_3}^{r_m} \lambda_{j_3}(\langle \mathbf{A}_r^*\mathbf{A}, M^*\rangle x_{j_3}^{t-1})^{\otimes l}$$

which again is a weighted sum of rank-1 symmetric tensors, thus being symmetric. This shows that $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta\nabla h^l(\mathbf{w}_{t-1})$ is also symmetric, concluding the induction step, thereby proving the claim.

Next, we show the breakdown of tensors along the GD trajectory

Lemma 17 *The GD trajectory of (4.7) $\{\mathbf{w}_t\}_{t=0}^\infty$ admits the following breakdown for an arbitrary t:*

$$\mathbf{w}_{t+1} = \langle \mathbf{Z}_t, \mathbf{w}_0\rangle - \mathbf{E}_t := \tilde{\mathbf{w}}_t - \mathbf{E}_t \tag{4.38}$$

where

$$\begin{aligned}
\mathbf{Z}_t &:= (\mathcal{I} + \eta\langle (\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)^t \\
\mathbf{E}_t &:= \sum_{i=1}^{t} (\mathcal{I} + \eta\langle (\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)^{t-i}\hat{\mathbf{E}}_i \\
\hat{\mathbf{E}}_i &:= \eta\langle\langle (\mathbf{A}_r^l)^*\mathbf{A}^l, \langle \mathbf{P}(\mathbf{w}_{i-1}), \mathbf{P}(\mathbf{w}_{i-1})\rangle_{2*[l]}\rangle, \mathbf{w}_{i-1}\rangle_{2*[l]}
\end{aligned}$$

and where $(\mathbf{A}_r^l)^*\mathbf{A}^l := \langle (\mathbf{A}_r)^{\otimes l}, \mathbf{A}^{\otimes l}\rangle_{3,6,\ldots,3l} \in \mathbb{R}^{[nr\times nr\times n\times n]\circ l}$ and $\mathcal{I}$ is the identity operator.

Proof 14 (Proof of Lemma 17) *For this proof, we will proceed by induction. For $t = 1$, we have that*

$$\begin{aligned}
\mathbf{w}_1 &= (\mathcal{I} + \eta\langle (\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l} - \langle \mathbf{P}(\mathbf{w}_0), \mathbf{P}(\mathbf{w}_0)\rangle\rangle)\mathbf{w}_0 \\
&= (\mathcal{I} + \eta\langle (\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)\mathbf{w}_0 - \eta\langle (\mathbf{A}_r^l)^*\mathbf{A}^l, \langle \mathbf{P}(\mathbf{w}_0), \mathbf{P}(\mathbf{w}_0)\rangle\rangle\mathbf{w}_0 \\
&= \langle \mathbf{Z}_1, \mathbf{w}_0\rangle - \mathbf{E}_1
\end{aligned}$$

Then, we move on to the induction step, while first assuming that it holds for some

t. One can write

$$\begin{aligned}
\mathbf{w}_{t+1} &= (\mathcal{I} + \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l} - \langle\mathbf{P}(\mathbf{w}_t), \mathbf{P}(\mathbf{w}_t)\rangle\rangle)\mathbf{w}_t \\
&= (\mathcal{I} + \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)\mathbf{w}_t - \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\mathbf{P}(\mathbf{w}_t), \mathbf{P}(\mathbf{w}_t)\rangle\rangle\mathbf{w}_t \\
&= (\mathcal{I} + \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)\mathbf{w}_t - \hat{\mathbf{E}}_{t+1} \\
&= (\mathcal{I} + \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)\left(\tilde{\mathbf{w}}_t - \sum_{i=1}^{t}(\mathcal{I} + \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)^{t-i}\hat{\mathbf{E}}_i\right) - \hat{\mathbf{E}}_{t+1} \\
&= \tilde{\mathbf{w}}_{t+1} - \sum_{i=1}^{t}(\mathcal{I} + \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)^{t+1-i}\hat{\mathbf{E}}_i - \hat{\mathbf{E}}_{t+1} \\
&= \tilde{\mathbf{w}}_{t+1} - \sum_{i=1}^{t+1}(\mathcal{I} + \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)^{t+1-i}\hat{\mathbf{E}}_i \\
&= \tilde{\mathbf{w}}_{t+1} - \mathbf{E}_t
\end{aligned}$$

Following the second step in the main outline, we aim to bound the spectral norm of $\mathbf{E}_t$, via the next lemma.

Lemma 18 *Given a tensor $\mathbf{E}_t$ defined in Lemma 17, assume that $\mathbf{w}_0 = \epsilon x_0^{\otimes l}$, where $\epsilon \in \mathbb{R}$ is the initialization scale. For every $t \leq t_s$,*

$$\|\mathbf{E}_t\|_S \leq \frac{8}{r_U^l \sigma_1(U)^l}\epsilon^3 (nL_s)^{l/2}(1 + \tilde{\eta}\sigma_1(U)^l)^{3t}\|x_0^{\otimes l}\|_*^3 \tag{4.39}$$

with

$$t_s = \lfloor \frac{\ln\left(\frac{\sigma_1^l(U) r_U^l}{8r^l L_s^{l/2}\|x_0^{\otimes l}\|_*^3}\frac{|x_0^\top v_1|^l}{n^{l/2}}\right) - 2\ln(\epsilon)}{2\ln(1 + \tilde{\eta}\sigma_1^l(U))}\rfloor \tag{4.40}$$

where $U = \langle \mathbf{A}_r^* \mathbf{A}, M^* \rangle \in \mathbb{R}^{nr \times nr}$, r_U being the rank of U, and $\tilde{\eta} = r_U^l \eta$. $\sigma_1(U)$ denotes the largest singular value of U, and v_1 being its associated singular vector.

Proof 15 (Proof of Lemma 18) *From Lemma 14 and the definition in Lemma 17, it is apparent that*

$$\|\mathbf{E}_t\|_S \leq \sum_{i=1}^{t} \|(\mathcal{I} + \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)^{t-i}\|_S \|\hat{\mathbf{E}}_i\|_* \tag{4.41}$$

We proceed to derive upper bounds on the norm terms separately, and then combine them together later. We first deal with $\|\hat{\mathbf{E}}_i\|_*$. By Lemma 14, we have that

$$\|\hat{\mathbf{E}}_i\|_* \leq \eta\|\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\mathbf{P}(\mathbf{w}_{i-1}), \mathbf{P}(\mathbf{w}_{i-1})\rangle\rangle\|_*\|\mathbf{w}_{i-1}\|_*$$

Now, assume that $\mathbf{w}_{i-1}$ *admits the following breakdown*

$$\mathbf{w}_{i-1} = \sum_{j=1}^{r_{i-1}} \lambda_j (x_j^{i-1})^{\otimes l}, \quad x_j^{i-1} \in \mathbb{R}^{nr},\ \|x_j^{i-1}\|_2 = 1 \tag{4.42}$$

where $\|\mathbf{w}_{i-1}\|_* = \sum_j |\lambda_j|$*. Therefore,*

$$\langle \mathbf{P}(\mathbf{w}_{i-1}), \mathbf{P}(\mathbf{w}_{i-1})\rangle = \sum_{j_1,j_2}^{r_{i-1},r_{i-1}} \lambda_{j_1}\lambda_{j_2} \langle \mathbf{P}((x_{j_1}^{i-1})^{\otimes l}), \mathbf{P}((x_{j_2}^{i-1})^{\otimes l})\rangle,$$

leading to

$$\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}(\mathbf{w}_{i-1}), \mathbf{P}(\mathbf{w}_{i-1})\rangle\rangle = \sum_{j_1,j_2}^{r_{i-1},r_{i-1}} \lambda_{j_1}\lambda_{j_2} \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}((x_{j_1}^{i-1})^{\otimes l}), \mathbf{P}((x_{j_2}^{i-1})^{\otimes l})\rangle\rangle.$$

For given indices j_1, j_2 *index, it follows from Lemma 4.2.1 that*

$$\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}((x_{j_1}^{i-1})^{\otimes l}), \mathbf{P}((x_{j_2}^{i-1})^{\otimes l})\rangle\rangle = (\langle \mathbf{A}_r^* \mathbf{A}, \operatorname{mat}(x_{j_1}^{i-1}) \operatorname{mat}(x_{j_2}^{i-1})^\top\rangle)^{\otimes l}$$

Now, according to the definition of $\mathbf{A}_r := I_r \oslash_{1,2} \mathbf{A}$*, where* $\oslash$ *denotes the kronecker product (a reshaped tensor vector product, where the subscript denotes the dimension with which kronecker product is applied with respect to* $\mathbf{A}$*), we know that*

$$\langle \mathbf{A}_r^* \mathbf{A}, \operatorname{mat}(x_{j_1}^{i-1}) \operatorname{mat}(x_{j_2}^{i-1})^\top\rangle = I_r \oslash \langle \mathbf{A}^* \mathbf{A}, \operatorname{mat}(x_{j_1}^{i-1}) \operatorname{mat}(x_{j_2}^{i-1})^\top\rangle$$

Hence, the eigenvalues of the LHS are just r *copies of that of the RHS [67]. This further implies*

$$\begin{aligned}
\|\langle \mathbf{A}_r^* \mathbf{A}, \operatorname{mat}(x_{j_1}^{i-1}) \operatorname{mat}(x_{j_2}^{i-1})^\top\rangle\|_* &= r\|\langle \mathbf{A}^* \mathbf{A}, \operatorname{mat}(x_{j_1}^{i-1}) \operatorname{mat}(x_{j_2}^{i-1})^\top\rangle\|_* \\
&\le r\sqrt{n}\|\langle \mathbf{A}^* \mathbf{A}, \operatorname{mat}(x_{j_1}^{i-1}) \operatorname{mat}(x_{j_2}^{i-1})^\top\rangle\|_F \\
&\le r\sqrt{nL_s}\| \operatorname{mat}(x_{j_1}^{i-1}) \operatorname{mat}(x_{j_2}^{i-1})^\top\|_F \\
&= r\sqrt{nL_s}
\end{aligned}$$

where the second last inequality follows from the RSS property, and the last equality follows from (4.42)*. Next, we apply Lemma 14 again with*

$$\|\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}((x_{j_1}^{i-1})^{\otimes l}), \mathbf{P}((x_{j_2}^{i-1})^{\otimes l})\|_* \le (\|\langle \mathbf{A}_r^* \mathbf{A}, \operatorname{mat}(x_{j_1}^{i-1}) \operatorname{mat}(x_{j_2}^{i-1})^\top\rangle\|_*)^l \le r^l (nL_s)^{l/2}$$

which leads to

$$\begin{aligned}
&\|\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}(\mathbf{w}_{i-1}), \mathbf{P}(\mathbf{w}_{i-1})\rangle\rangle\|_* \\
\le &\sum_{j_1,j_2}^{r_{i-1},r_{i-1}} |\lambda_{j_1}||\lambda_{j_2}|\|\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}((x_{j_1}^{i-1})^{\otimes l}), \mathbf{P}((x_{j_2}^{i-1})^{\otimes l})\|_* \\
\le &r^l (nL_s)^{l/2} \sum_{j_1,j_2}^{r_{i-1},r_{i-1}} |\lambda_{j_1}||\lambda_{j_2}| = r^l (nL_s)^{l/2} \|\mathbf{w}_{i-1}\|_*^2
\end{aligned}$$

This directly gives

$$\|\hat{\mathbf{E}}_i\|_* \leq \eta(r^2 nL_s)^{l/2}\|\mathbf{w}_{i-1}\|_*^3$$

Since our goal is to bound $\|\mathbf{E}_t\|_S$*, we focus on* $\|(\mathcal{I}+\eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle)^{t-i}\|_S$*. Using binomial formula, we obtain that*

$$(\mathcal{I}+\eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle)^{t-i} = \sum_{k=0}^{t-i}\binom{t-i}{k}\eta^k(\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle)^k$$

where $\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle \in \mathbb{R}^{nr\circ 2l}$*, and* $(\cdot)^k$ *just denotes repeated multiplications along the even dimensions of the tensor, as explained in the disclaimer. To upper-bound the spectral norm of* $(\mathcal{I}+\eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle)^{t-i}$*, it is necessary to upper-bound the spectral norm of* $(\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle)^k$*. To do so, we use Lemma 15 to reformulate*

$$\|\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle^k\|_S = \sup_{\|\mathbf{v}\|_*\leq 1}|\langle\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle^k,\mathbf{v}\rangle|$$

Assume that the above supremum is achieved at $\mathbf{v}^*$*, with nuclear norm decomposition of*

$$\mathbf{v}^* = \sum_{j_v=1}^{r_v}\lambda_{j_v}x_{j_v,1}\otimes\cdots\otimes x_{j_v,2l}, \quad x_{j_v,p}\in\mathbb{R}^{nr},\ \|x_{j_v,p}\|_2 = 1\ \forall p\in[2l]$$

with $\sum_{j_v}|\lambda_{j_v}| = \|\mathbf{v}^*\|_* \leq 1$*. Note that this decomposition is due to the fact that* $\mathbf{v}$ *is not necessarily symmetric. Again, by Lemma 4.2.1,*

$$\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle^k = \left[(\langle\mathbf{A}_r^*\mathbf{A},M^*\rangle)^k\right]^{\otimes l},$$

directly leading to

$$\|\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle^k\|_S = \sum_{j_v=1}^{r_v}|\lambda_{j_v}\prod_{p=0}^{l-1}x_{j_v,p*2}^\top\langle\mathbf{A}_r^*\mathbf{A},M^*\rangle^k x_{j_v,p*2+1}|$$

Since

$$x_{j_v,p*2}^\top\langle\mathbf{A}_r^*\mathbf{A},M^*\rangle^k x_{j_v,p*2+1} \leq \sigma_1^k(U)$$

this means that

$$\|\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle^k\|_S = (\sigma_1^k(U))^l\sum_{j_v=1}^{r_v}|\lambda_{j_v}| \leq \sigma_1^{kl}(U)$$

Going back to $(\mathcal{I}+\eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle)^{t-i}$*,*

$$\begin{aligned}\|(\mathcal{I}+\eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle)^{t-i}\|_S &\leq \sum_{k=0}^{t-i}\binom{t-i}{k}\eta^k\|(\langle(\mathbf{A}_r^l)^*\mathbf{A}^l,(M^*)^{\otimes l}\rangle)^k\|_S \\ &\leq \sum_{k=0}^{t-i}\binom{t-i}{k}\eta^k\sigma_1^{kl}(U) = (1+\eta\sigma_1^l(U))^{t-i}.\end{aligned}$$

Before further upper-bounding $\|\mathbf{E}_t\|_S$*, we define* t_s *in such a way that*

$$\|\tilde{\mathbf{w}}_t - \mathbf{w}_t\|_* \leq \|\tilde{\mathbf{w}}_t\|_*, \quad \forall t \leq t_s \tag{4.43}$$

where $\tilde{\mathbf{w}}_t$ *is defined in* (4.38)*. We will later justify the existence of* t_s *and derive a lower bound. If the above inequality holds true, we also have*

$$\|\mathbf{w}_t\|_* \leq \|\tilde{\mathbf{w}}_t\|_* + \|\tilde{\mathbf{w}}_t - \mathbf{w}_t\|_* \leq 2\|\tilde{\mathbf{w}}_t\|_*.$$

Recall the binomial formula again and decompose $\tilde{\mathbf{w}}_t$ *into*

$$\tilde{\mathbf{w}}_t = \sum_{k=0}^{t} \binom{t}{k} \eta^k \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, (M^*)^{\otimes l} \rangle^k \mathbf{w}_0 \tag{4.44}$$

Therefore, it follows from Lemma 14 that,

$$\|\tilde{\mathbf{w}}_{i-1}\|_* \leq \left(\sum_{k=0}^{i-1} \binom{i-1}{k} \eta^k \|\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, (M^*)^{\otimes l} \rangle^k\|_* \right) \|\mathbf{w}_0\|_* \tag{4.45}$$

for all $i \leq t$*. With the repeated application of Lemma 14, we have*

$$\|\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, (M^*)^{\otimes l} \rangle^k\|_* \leq (\|U\|_*)^{kl} \leq \left(r_U^l \sigma_1^l(U)\right)^k$$

Therefore, substituting back into (4.45) *gives*

$$\|\tilde{\mathbf{w}}_{i-1}\|_* \leq \left(\sum_{k=0}^{t} \binom{t}{k} \eta^k \left(r^l r_U^l \sigma_1^l(U)\right)^k \right) \|\mathbf{w}_0\|_* = (1 + \tilde{\eta}\sigma_1^l(U))^{i-1} \|\mathbf{w}_0\|_*$$

Next, plugging the above preparatory results into (4.41)*, we have that*

$$\begin{aligned}
\|\mathbf{E}_t\|_S &\leq \sum_{i=1}^{t}(1+\eta\sigma_1^l(U))^{t-i}\eta(r^2nL_s)^{l/2}\|\mathbf{w}_{i-1}\|_*^3 \\
&\leq \sum_{i=1}^{t}(1+\eta\sigma_1^l(U))^{t-i}\eta(r^2nL_s)^{l/2}8\|\tilde{\mathbf{w}}_{i-1}\|_*^3 \\
&\leq 8\sum_{i=1}^{t}(1+\eta\sigma_1^l(U))^{t-i}\eta(r^2nL_s)^{l/2}(1+\tilde{\eta}\sigma_1^l(U))^{3i-3}\|\mathbf{w}_0\|_*^3 \\
&\leq 8\epsilon^3\eta(r^2nL_s)^{l/2}\sum_{i=1}^{t}(1+\tilde{\eta}\sigma_1^l(U))^{t-i}(1+\tilde{\eta}\sigma_1^l(U))^{3i-3} \\
&= 8\epsilon^3\|x_0^{\otimes l}\|_*^3\eta(r^2nL_s)^{l/2}(1+\tilde{\eta}\sigma_1^l(U))^{t-1}\sum_{i=1}^{t}(1+\tilde{\eta}\sigma_1^l(U))^{2i-2} \\
&= 8\epsilon^3\|x_0^{\otimes l}\|_*^3\eta(r^2nL_s)^{l/2}(1+\tilde{\eta}\sigma_1^l(U))^{t-1}\frac{(1+\tilde{\eta}\sigma_1^l(U))^{2t}-1}{(1+\tilde{\eta}\sigma_1^l(U))^2-1} \quad \textit{(geometric sum)} \\
&\leq 8\epsilon^3\|x_0^{\otimes l}\|_*^3\eta(r^2nL_s)^{l/2}(1+\tilde{\eta}\sigma_1^l(U))^{t-1}(1+\tilde{\eta}\sigma_1^l(U))^{2t} \\
&\leq \frac{8\eta}{\tilde{\eta}\sigma_1^l(U)}\epsilon^3(r^2nL_s)^{l/2}(1+\tilde{\eta}\sigma_1^l(U))^{3t}\|x_0^{\otimes l}\|_*^3 \\
&= \frac{r^l8}{r_U^l\sigma_1^l(U)}\epsilon^3(nL_s)^{l/2}(1+\tilde{\eta}\sigma_1^l(U))^{3t}\|x_0^{\otimes l}\|_*^3
\end{aligned}$$

proving the original claim of this lemma (4.39)*. Now, we give a lower bound on* t_s*. By recalling the breakdown* (4.44)*, we have*

$$\begin{aligned}
\|\tilde{\mathbf{w}}_t\|_* \geq \|\tilde{\mathbf{w}}_t\|_S &\geq \langle\tilde{\mathbf{w}}_t, v_1^{\otimes l}\rangle \\
&= \epsilon\sum_{k=0}^{t}\binom{t}{k}\eta^k\left[|v_1^\top\langle\mathbf{A}_r^*\mathbf{A}, M^*\rangle^k x_0|\right]^l \\
&= \epsilon\sum_{k=0}^{t}\binom{t}{k}\eta^k\left[|v_1^\top U^k x_0|\right]^l \\
&= \epsilon\sum_{k=0}^{t}\binom{t}{k}\eta^k(|\sigma_1^k(U)v_1^\top x_0|)^l = \epsilon|v_1^\top x_0|^l(1+\eta\sigma_1^l(U))^t
\end{aligned} \tag{4.46}$$

with v_1 *being the first singular vector of* $I_r \oslash U$*. Since the sensing matrices are assumed to be symmetric,* U *is also symmetric, hence the singular vectors of* U^k *coincide with those of* U*. By* (4.39)*, we also know*

$$\frac{\|\tilde{\mathbf{w}}_t - \mathbf{w}_t\|_*}{\|\tilde{\mathbf{w}}_t\|_*} \leq \frac{r^l8}{r_U^l\sigma_1^l(U)}\epsilon^2\|x_0^{\otimes l}\|_*^3\frac{n^{l/2}}{(v_1^\top x_0)^l}L_s^{l/2}\frac{(1+\tilde{\eta}\sigma_1^l(U))^{3t}}{(1+\eta\sigma_1^l(U))^t}$$

Therefore, for (4.43) *to hold true, we need the RHS of the above equation to be smaller than 1, meaning that*

$$3t\ln(1+\tilde{\eta}\sigma_1^l(U)) \le \ln\left(\frac{r_U^l\sigma_1^l(U)}{8r^l\epsilon^2 L_s^{l/2}\|x_0^{\otimes l}\|_*^3}\frac{(v_1^\top x_0)^l}{n^{l/2}}\right) + t\ln(1+\eta\sigma_1^l(U))$$

This further implies that for (4.43) *to hold, t should satisfy*

$$t < \frac{\ln\left(\frac{r_U^l\sigma_1^l(U)}{8r^l\epsilon^2 L_s^{l/2}\|x_0^{\otimes l}\|_*^3}\frac{(v_1^\top x_0)^l}{n^{l/2}}\right)}{3\ln(1+\tilde{\eta}\sigma_1^l(U)) - \ln(1+\eta\sigma_1^l(U))} < \frac{\ln\left(\frac{r_U^l\sigma_1^l(U)}{8r^l\epsilon^2 L_s^{l/2}\|x_0^{\otimes l}\|_*^3}\frac{(v_1^\top x_0)^l}{n^{l/2}}\right)}{2\ln(1+\tilde{\eta}\sigma_1^l(U))}$$

which after rearrangement gives (4.40)*.*

Now, we present the proof of Lemma 13.

Proof 16 (Proof of Lemma 13) *Using the tensor Weyl's inequality (Lemma 16), we have that*

$$\lambda_2^v(\mathbf{w}_t) \le \lambda_2^v(\tilde{\mathbf{w}}_t) + \|\mathbf{E}_t\|_S \tag{4.47}$$

$$\lambda_1^v(\mathbf{w}_t) \ge \lambda_1^v(\tilde{\mathbf{w}}_t) - \|\mathbf{E}_t\|_S \tag{4.48}$$

The only remaining part of the proof is the characterization of $\lambda_1^v(\tilde{\mathbf{w}}_t)$ and $\lambda_2^v(\tilde{\mathbf{w}}_t)$. The first term is easy because we already have the characterization from the proof of Lemma 18, with (4.46) *giving rise to*

$$\|\tilde{\mathbf{w}}_t\|_S \ge \epsilon|v_1^\top x_0|^l(1+\eta\sigma_1^l(U))^t$$

Also, by the definition of v-eigenvalues and (4.44)*, we have that*

$$\begin{aligned}
\lambda_2^v(\tilde{\mathbf{w}}_t) &= \max_{\substack{V\\ \dim(V)=2}} \min_{\substack{v\in V\\ \|v\|_2=1}} \epsilon\sum_{k=0}^t \binom{t}{k}\eta^k\left[|v^\top\langle\mathbf{A}_r^*\mathbf{A}, M^*\rangle^k x_0|\right]^l \\
&= \epsilon\|x_0\|_2^l \max_{\substack{V\\ \dim(V)=2}} \min_{\substack{v\in V\\ \|v\|_2=1}} \sum_{k=0}^t \binom{t}{k}\eta^k|v^\top U^k\frac{x_0}{\|x_0\|_2}|^l \\
&\le \epsilon\|x_0\|_2^l \max_{\substack{V\\ \dim(V)=2}} \min_{\substack{v\in V\\ \|v\|_2=1}} \sum_{k=0}^t \binom{t}{k}\eta^k|v^\top U^k v|^l \\
&= \epsilon\|x_0\|_2^l\sum_{k=0}^t \binom{t}{k}\eta^k|v_2^\top U^k v_2|^l \\
&= \epsilon\|x_0\|_2^l\sum_{k=0}^t \binom{t}{k}\eta^k|\sigma_2^k(U)|^l \\
&= \epsilon\|x_0\|_2^l(1+\eta\sigma_2^l(U))^t
\end{aligned}$$

where v_2 is the singular vector associated with $\sigma_2^k(U)$ $\forall k\in[t]$. Finally, combining the above equations yields (4.14) *after rearrangements.*

Next, we present a supporting lemma which explains that Gaussian concentration is suited for our purpose.

Lemma 19 *Let $x_0 = v_1 + g \in \mathbb{R}^{nr}$, where g is a vector with each entry being i.i.d sampled from Gaussian distribution $\mathcal{N}(0, \rho)$. For some universal constant C, the follwoing inequalities hold:*

$$\mathbb{P}\left[|v_1^\top x_0|^l \geq (1 - \mathcal{O}(\sqrt{\rho}))^l\right] \geq 1 - 2\exp(-C/\rho),$$
$$\mathbb{P}\left[\|x_0\|_2^l \leq (\sqrt{1+\rho^2 nr} + \mathcal{O}(\rho^{3/2}))^l\right] \geq 1 - 2\exp(-C/\rho)$$

Proof 17 (Proof of Lemma 19) *We know that*

$$|v_1^\top x_0| = |1 + v_1^\top g| \geq 1 - |v_1^\top x_0|$$

Theorem 2.6.3 of [79] (general Hoeffding's) gives that with probability at least $1 - 2\exp(-t^2/\rho^2)$,

$$|v^\top g| \leq t \quad \forall \|v\|_2 = 1$$

which leads to the first concentration bound after substituting $t = \mathcal{O}(\sqrt{\rho})$ with some constant c_1. Then, Theorem 3.1.1 in [79] gives

$$\mathbb{P}\left[|\|x_0\|_2 - \sqrt{1+\rho^2 nr}| \leq t\right] \geq 1 - 2\exp(-c_2 t^2/\rho^4)$$

for $g \sim \mathcal{N}(0, \rho I_{nr})$ and some constant c_2. This is because $\mathbb{E}[\|x_0\|_2^2] = 1 + \rho^2 nr$. Substituting $t = \mathcal{O}(\rho^{3/2})$ yields that

$$\mathbb{P}\left[\|x_0\|_2 \leq \sqrt{1+\rho^2 nr} + \mathcal{O}(\rho^{3/2})\right] \geq 1 - 2\exp(-c_2/\rho)$$

which results in the second bound. Now, we choose $C = \min\{c_1, c_2\}$.

Then, we prove our main theorem of this section.

Proof 18 (Proof of Theorem 6) *First, set $2\zeta = \kappa$, implying that $\zeta < 1/2$. We aim to derive sufficient conditions for the following inequalities to hold:*

$$\lambda_2^v(\tilde{\mathbf{w}}_t) \leq \frac{\zeta}{2}\lambda_1^v(\tilde{\mathbf{w}}_t), \tag{4.49}$$
$$\|\mathbf{E}_t\|_s \leq \frac{\zeta}{2}\lambda_1^v(\tilde{\mathbf{w}}_t) \tag{4.50}$$

By recalling Lemma 13, a sufficient condition for (4.49) is that

$$\epsilon\|x_0\|_2^l(1 + \eta\sigma_2^l(U))^t \leq \frac{\zeta}{2}\epsilon|v_1^\top x_0|^l(1 + \eta\sigma_1^l(U))^t$$

implying that

$$\frac{2\|x_0\|_2^l}{\zeta|v_1^\top x_0|^l} \leq \left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^t$$

which after rearrangements gives $t \geq t(\zeta, l)$*, as defined in* (4.17)*. Then, we obtain a sufficient condition for* (4.50)*, which by Lemma 18 is*

$$\frac{8r^l}{r_U^l\sigma_1(U)^l}\epsilon^3(nL_s)^{l/2}(1+\tilde{\eta}\sigma_1(U)^l)^{3t}\|x_0^{\otimes l}\|_*^3 \leq \frac{2}{\zeta}\epsilon|v_1^\top x_0|^l(1+\eta\sigma_1^l(U))^t \tag{4.51}$$

contingent on the fact that $t \leq t_s$*. Therefore, before going further, we need to verify that* $t(\zeta, l) \leq t_s$ *for some small enough* ϵ*.* (4.40) *implies that a sufficient condition is*

$$\ln\left(\frac{2\|x_0\|_2^l}{\zeta|v_1^\top x_0|^l}\right)\ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1} \leq \frac{\ln\left(\frac{\sigma_1^l(U)r_U^l}{8r^lL_s^{l/2}\|x_0^{\otimes l}\|_*^3\epsilon^2}\frac{|x_0^\top v_1|^l}{n^{l/2}}\right)}{2\ln(1+\tilde{\eta}\sigma_1^l(U))}$$

Additionally, by leveraging the identity $x/(1+x) \leq \ln(1+x) \leq x$*, we derive the following identity*

$$\frac{\ln(1+\tilde{\eta}\sigma_1^l(U))}{\ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1}} \leq \frac{r_U^l(1+\eta\sigma_1^l(U))}{1-(\sigma_2(U)/\sigma_1(U))^l} := \Xi \tag{4.52}$$

Hence,

$$2\ln\left(\frac{2\|\mathbf{w}_0\|_2^l}{\zeta|v_1^\top x_0|^l}\right)\Xi \leq \ln\left(\frac{\sigma_1^l(U)r_U^l}{8r^lL_s^{l/2}\|x_0^{\otimes l}\|_*^3\epsilon^2}\frac{|x_0^\top v_1|^l}{n^{l/2}}\right)$$

and after rearrangement gives

$$\epsilon^2 \leq \frac{\sigma_1^l(U)r_U^l}{8(r^2nL_s)^{l/2}}\frac{|x_0^\top v_1|^l}{\|x_0^{\otimes l}\|_*^3}\left(\frac{2\|x_0\|_2^l}{\zeta|v_1^\top x_0|^l}\right)^{-\Xi} \tag{4.53}$$

Notice that all of the above terms are independent of ϵ*, and are positive. Therefore, a small enough* ϵ *exists. Also notice that a smaller step-size* η *will yield a loser bound on* ϵ *through the dependence of* Ξ*. Now, consider* (4.51) *again. Since* T *is finite, a sufficient condition for* (4.51) *is*

$$\epsilon^2 \leq \zeta\frac{r_U^l\sigma_1(U)^l}{16(r^2nL_s)^{l/2}}\frac{|v_1^\top x_0|^l}{\|x_0^{\otimes l}\|_*^3}\left(\frac{1+\eta\sigma_1^l(U)}{(1+\tilde{\eta}\sigma_1(U)^l)^3}\right)^T \tag{4.54}$$

which can again be achieved by setting a small enough ϵ*, since all other terms are positive and not dependent on it. In summary, if we choose a small constant* ϵ *satisfying both* (4.53) *and* (4.54)*, and if* $t_s \geq t_T$ *(which again can be achieved via a* *sufficiently small ϵ), it is already sufficient for both* (4.49) *and* (4.50) *to hold, thereby giving:*

$$\frac{\lambda_2^v(\tilde{\mathbf{w}}_t) + \|\mathbf{E}_t\|_S}{\lambda_1^v(\tilde{\mathbf{w}}_t)} \leq \zeta$$

If $\zeta < 1/2$, this further implies

$$\lambda_1^v(\tilde{\mathbf{w}}_t) > 2\lambda_2^v(\tilde{\mathbf{w}}_t) + 2\|\mathbf{E}_t\|_S \implies \|\mathbf{E}_t\|_S \leq \frac{1}{2}\lambda_1^v(\tilde{\mathbf{w}}_t) - \lambda_2^v(\tilde{\mathbf{w}}_t) \leq \frac{1}{2}\lambda_1^v(\tilde{\mathbf{w}}_t)$$

As a result,

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \leq \frac{\lambda_2^v(\tilde{\mathbf{w}}_t) + \|\mathbf{E}_t\|_S}{\lambda_1^v(\tilde{\mathbf{w}}_t) - \|\mathbf{E}_t\|_S} \leq \frac{\zeta\lambda_1^v(\tilde{\mathbf{w}}_t)}{\lambda_1^v(\tilde{\mathbf{w}}_t)/2} = 2\zeta$$

which proves (4.16)*.*

Theorem 6 can also be improved via Lemma 19 as stated below.

Corollary 1 (Corollary to Theorem 6) *Consider the optimization problem and the GD trajectory given in Theorem 6. If additionally $x_0 = v_1 + g \in \mathbb{R}^{nr}$ and $g \sim \mathcal{N}(0, \rho I_{nr})$, then*

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \leq \kappa \quad \textit{for} \quad t \asymp \ln\left(\frac{1}{\kappa}\right) \ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1} \tag{4.55}$$

provided that

$$\epsilon \asymp \sqrt{\kappa/2}\frac{(\sigma_1(U)r_U)^{l/2}}{4(r^2nL_s)^{l/4}}(\frac{\kappa}{4})^{3\Xi/2}, \quad \textit{where } \Xi := \frac{r_U^l(1+\eta\sigma_1^l(U))}{1-(\sigma_2(U)/\sigma_1(U))^l} \tag{4.56}$$

with probability at least $1 - 2\exp(-C/\rho)$ for some universal constant C as $\rho \to 0$, where $\sigma_1(U)$ and $\sigma_2(U)$ are the first two singular values of $U = \langle \mathbf{A}_r, b\rangle_3$, with v_1 being the associated singular vector of $\sigma_1(U)$ ($\asymp$ denotes "asymptotic to", meaning that the two terms of both sides of this symbol are of the same order of magnitude).

Proof 19 (Proof of Corollary 1) *The proof is similar to that of Theorem 6 (note $\zeta = \kappa/2$), and therefore we only highlight the difference. We know that* (4.26) *holds true if*

$$t \geq \ln\left(\frac{2\|x_0\|_2^l}{\zeta|v_1^\top x_0|^l}\right) \ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1},$$
$$\epsilon^2 \leq \frac{\sigma_1^l(U)r_U^l}{8(r^2nL_s)^{l/2}} \frac{|x_0^\top v_1|^l}{\|x_0^{\otimes l}\|_*^3} \left(\frac{2\|x_0\|_2^l}{\zeta|v_1^\top x_0|^l}\right)^{-\Xi}$$

It results from Lemma 19 that for our choice of initialization, we have that

$$\|x_0\|_2^l \asymp \|v_1^\top x_0|^l \asymp 1$$

with probability at least $1 - 2\exp(-C/\rho)$*. Thus, as long as*

$$t \asymp \ln\left(\frac{2}{\zeta}\right) \ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1} := t_*, \tag{4.57}$$

$$\epsilon \asymp \frac{(\sigma_1(U) r_U)^{l/2}}{2\sqrt{2}(r^2 n L_s)^{l/4}} \left(\frac{2}{\zeta}\right)^{-\Xi/2} \tag{4.58}$$

(4.26) *will hold with high probability. Next, in order for* (4.50) *to hold for* $t \asymp t_*$*, we know that*

$$\epsilon^2 \leq \zeta \frac{r_U^l \sigma_1(U)^l}{16(r^2 n L_s)^{l/2}} \frac{|v_1^\top x_0|^l}{\|x_0^{\otimes l}\|_*^3} \left(\frac{1+\eta\sigma_1^l(U)}{(1+\tilde{\eta}\sigma_1(U)^l)^3}\right)^{t_*}$$

Via the same order of magnitude argument, we know that the following condition is sufficient for (4.50)*:*

$$\epsilon \asymp \sqrt{\zeta} \frac{(\sigma_1(U) r_U)^{l/2}}{4(r^2 n L_s)^{l/4}} \left(\frac{1+\eta\sigma_1^l(U)}{(1+\tilde{\eta}\sigma_1(U)^l)^3}\right)^{t_*/2}$$

Now,

$$\begin{aligned}
\left(\frac{1+\eta\sigma_1^l(U)}{(1+\tilde{\eta}\sigma_1(U)^l)^3}\right)^{t_*/2} &\geq \left[\frac{1}{(1+\tilde{\eta}\sigma_1(U)^l)^3}\right]^{t_*/2} \\
&= \exp\left(-\frac{3t_*}{2}\ln(1+\tilde{\eta}\sigma_1(U)^l)\right) \\
&= \exp\left(-\frac{3}{2}\ln(\frac{2}{\zeta})\frac{\ln(1+\tilde{\eta}\sigma_1(U)^l)}{\ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)}\right) \\
&\geq \exp\left(-\frac{3}{2}\ln(\frac{2}{\zeta})\Xi\right) = (\frac{\zeta}{2})^{3\Xi/2}
\end{aligned}$$

where the second equality follows from the substitution of t_**, and the last inequality follows from* (4.52)*. As a result,*

$$\epsilon \asymp \sqrt{\zeta} \frac{(\sigma_1(U) r_U)^{l/2}}{4(r^2 n L_s)^{l/4}} (\frac{\zeta}{2})^{3\Xi/2} \tag{4.59}$$

Therefore, taking the minimum of (4.58) *and* (4.59)*, we know that*

$$\epsilon \asymp \sqrt{\zeta} \frac{(\sigma_1(U) r_U)^{l/2}}{(r^2 n L_s)^{l/4}} (\frac{\zeta}{2})^{3\Xi/2} \tag{4.60}$$

is sufficient for (4.49) *and* (4.50)*, leading to* (4.55) *via the same steps in the proof of Theorem 6.*

4.B.3 Additional Details for Properties of Approximate Rank-1 Tensors

We start with the proof of Proposition 2.

Proof 20 (Proof of Proposition 2) *Given a symmetric tensor* $\mathbf{w}$*, it can be decomposed as*

$$\mathbf{w} = \sum_{i=1}^{r_w} \lambda_i x_i^{\otimes l}$$

where r_w *is* $\mathbf{w}$*'s symmetric rank. Now, consider the vector* $w_s \in \mathbb{R}^n$ *that attains the spectral norm, meaning that* $\langle \mathbf{w}, w_s^{\otimes l} \rangle = \lambda_1^v(\mathbf{w})$*. One can decompose each* $x_i^{\otimes l}$ *into a parallel component and an orthogonal component. To be more specific,*

$$x_i = x_i^s + x_i^{\perp} \implies x_i^{\otimes l} = (x_i^s)^{\otimes l} + \sum_{j=1}^{2^l-1} \underbrace{x_i^{\perp} \otimes \cdots \otimes x_i^{\perp}}_{j} \otimes \underbrace{x_i^s \otimes \cdots \otimes x_i^s}_{l-j}$$

and it is apparent that the second term is orthogonal to $w_s^{\otimes l}$ *via Lemma 4.2.1. Therefore, we just organize all components* $w_s^{\otimes l}$ *together and all orthogonal components together. By definition, the parallel component has the magnitude* $\lambda_1^v(\mathbf{w})$*. Also, by the definition of v-eigenvalues,* $\|\mathbf{w}^{\dagger}\|_S \leq \lambda_2^v(\mathbf{w}_t)$ *since otherwise the dominant direction of* $\mathbf{w}^{\dagger}$ *will just become the second eigenvector of* $\mathbf{w}$*.*

We now provide the proof of Proposition 3.

Proof 21 (Proof of Proposition 3) *According to* (4.9a)*, the gradient of* (4.7) *with respect to* $\mathbf{w}$ *can be expressed as*

$$\nabla h^l(\mathbf{w}) = \langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{w}) \rangle_{2*[l]} - (M^*)^{\otimes l} \rangle, \mathbf{w} \rangle_{2*[l]} \tag{4.61}$$

where $(\mathbf{A}_r^l)^* \mathbf{A}^l$ *is defined in* (4.37)*. In light of* (4.18)*, one can write*

$$\begin{aligned}
&\langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{w}) \rangle_{2*[l]} = \langle \langle \mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l} \rangle_{2*[l]}, \mathbf{w} \otimes \mathbf{w} \rangle_{3,4,7,8,\ldots,4l-1,4l} \\
&= \underbrace{\langle \langle \mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l} \rangle, \mathbf{w}_\sigma \otimes \mathbf{w}_\sigma \rangle}_{\mathbf{a}_1} + 2 \underbrace{\langle \langle \mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l} \rangle, \mathbf{w}_\sigma \otimes \mathbf{w}^{\dagger} \rangle}_{\mathbf{a}_2} + \underbrace{\langle \langle \mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l} \rangle, \mathbf{w}^{\dagger} \otimes \mathbf{w}^{\dagger} \rangle}_{\mathbf{a}_3}
\end{aligned}$$

where $\mathbf{w}_\sigma = \lambda_1^v(\mathbf{w}) \hat{w}^{\otimes l}$*. Note that we have dropped the subscripts from the second line and henceforth for sake of simplicity. By using this logic,* (4.61) *can be written as*

$$\nabla h^l(\mathbf{w}) = \underbrace{\langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{a}_1 - (M^*)^{\otimes l} \rangle, \mathbf{w}_\sigma \rangle}_{\mathbf{h}_1} + \mathbf{h}_2$$

where

$$\begin{aligned}
\mathbf{h}_2 = &\langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \mathbf{a}_1 \rangle, \mathbf{w}^{\dagger} \rangle + \langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \mathbf{a}_2 \rangle, \mathbf{w}_\sigma \rangle + \langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \mathbf{a}_2 \rangle, \mathbf{w}^{\dagger} \rangle + \\
&\langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \mathbf{a}_3 \rangle, \mathbf{w}_\sigma \rangle + \langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \mathbf{a}_3 \rangle, \mathbf{w}^{\dagger} \rangle - \langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, (M^*)^{\otimes l} \rangle, \mathbf{w}^{\dagger} \rangle
\end{aligned}$$

The first term can be analyzed as

$$\langle\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \mathbf{a}_1\rangle, \mathbf{w}^\dagger\rangle = \langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \mathbf{w}_\sigma \otimes \mathbf{w}_\sigma \otimes \mathbf{w}^\dagger\rangle\rangle$$

and by Lemma 14, we have that

$$\begin{aligned}\|\langle\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \mathbf{a}_1\rangle, \mathbf{w}^\dagger\rangle\|_S &\leq \|\langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \mathbf{w}_\sigma \otimes \mathbf{w}_\sigma \otimes \mathbf{w}^\dagger\rangle\|_S \|(\mathbf{A}_r^l)^*\mathbf{A}^l\|_* \\ &= \|\langle\mathbf{P}(\mathbf{w}_\sigma), \mathbf{P}(\mathbf{w}_\sigma)\rangle \otimes \mathbf{w}^\dagger\|_S \|(\mathbf{A}_r^l)^*\mathbf{A}^l\|_* \\ &\leq \lambda_1^v(\mathbf{w})^2 \|\mathbf{w}^\dagger\|_S \|(\mathbf{A}_r^l)^*\mathbf{A}^l\|_* \leq \kappa\lambda_1^v(\mathbf{w})^3 r^l \|\mathbf{A}^*\mathbf{A}\|_*^l\end{aligned} \tag{4.62}$$

The second inequality follows form that for all $u_1 \in \mathbb{R}^n$ and $u_2 \in \mathbb{R}^{nr}$ such that $\|u_1\|_2 = 1$ and $\|u_2\|_2 = 1$:

$$\begin{aligned}\|\langle\mathbf{P}(\mathbf{w}_\sigma), \mathbf{P}(\mathbf{w}_\sigma)\rangle \otimes \mathbf{w}^\dagger\|_S &= \max_{u_1,u_2} \langle\langle\mathbf{P}(\mathbf{w}_\sigma), \mathbf{P}(\mathbf{w}_\sigma)\rangle \otimes \mathbf{w}^\dagger, u_1^{\otimes 2l} \otimes u_2^{\otimes l}\rangle \\ &\leq \lambda_1^v(\mathbf{w})^2 (u^\top \operatorname{mat}(\hat{x}) \operatorname{mat}(\hat{x})^\top u)^l \|\mathbf{w}^\dagger\|_S \\ &\leq \lambda_1^v(\mathbf{w})^2 \sigma_{\max}(\operatorname{mat}(\hat{x}))^{2l} \|\mathbf{w}^\dagger\|_S \\ &\leq \lambda_1^v(\mathbf{w})^2 \|\hat{x}\|_2^{2l} \|\mathbf{w}^\dagger\|_S \\ &= \lambda_1^v(\mathbf{w})^2 \|\mathbf{w}^\dagger\|_S\end{aligned}$$

Repeating this process leads to

$$\|\mathbf{h}_2\|_S \leq (3\kappa + 3\kappa^2 + \kappa^3 + \kappa\|M^*\|_F^2)\lambda_1^v(\mathbf{w})^3 r^l \|\mathbf{A}^*\mathbf{A}\|_*^l \tag{4.63}$$

Similarly, $\|\mathbf{h}_1\|_S = \mathcal{O}(\lambda_1^v(\mathbf{w})^3 r^l \|\mathbf{A}^*\mathbf{A}\|_*^l)$. Now, if we assume that $\mathbf{w}$ is an FOP of (4.7)*, it means that $\nabla h^l(\mathbf{w}) = 0$, further implying $\|\nabla h^l(\mathbf{w})\|_S = 0$, and by reverse triangle inequality,*

$$0 = \|\nabla h^l(\mathbf{w})\|_S \geq |\|\mathbf{h}_1\|_S - \|\mathbf{h}_2\|_S|$$

which means that $\|\mathbf{h}_1\|_S = \|\mathbf{h}_2\|_S$. Since there always exits a small enough κ such that $\|\mathbf{h}_2\|_S = c\|\mathbf{h}_1\|_S$ with $c < 1$, and therefore the only possibility that the above inequality holds true is that $\|\mathbf{h}_1\|_S = \|\mathbf{h}_2\|_S = 0$. This implies

$$\langle\mathbf{h}_1, u^{\otimes l}\rangle = (\langle\mathbf{A}, \operatorname{mat}(w_s)\operatorname{mat}(w_s)^\top - M^*\rangle^\top \langle\mathbf{A}, \operatorname{mat}(w_s)\operatorname{mat}(u)^\top\rangle)^l = 0 \quad \forall u \in \mathbb{R}^{nr}$$

which is equivalent to the FOP condition for (1.3)*, which is* (2.6)*, meaning that* $\operatorname{mat}(w_s) \in \mathbb{R}^{n\times r}$ *is an FOP of* (1.3)*. Note that we can always scale $\mathbf{A}$ and b together so that $\|\mathbf{A}^*\mathbf{A}\|_*^l$ can be normalized to 1.*

Finally, we prove the main result of this paper.

Proof 22 (Proof of Theorem 7) *We consider the SOP condition for* (4.7)*, which is* (4.9b) *for some rank-1 tensor Δ. We can express it as*

$$\begin{aligned}\nabla^2 h^l(\hat{\mathbf{w}})[\Delta, \Delta] =& 2\underbrace{\langle\nabla f^l(\langle\mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}})\rangle_{2*[l]}), \langle\mathbf{P}(\Delta), \mathbf{P}(\Delta)\rangle_{2*[l]}}_{\mathbf{a}_1(\hat{\mathbf{w}})} + \\ &\underbrace{\|\langle\mathbf{A}^{\otimes l}, \langle\mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\Delta)\rangle_{2*[l]} + \langle\mathbf{P}(\Delta), \mathbf{P}(\hat{\mathbf{w}})\rangle_{2*[l]}\rangle\|_F^2}_{\mathbf{a}_2(\hat{\mathbf{w}})}\end{aligned}$$

Let Δ *be defined identically to that in the proof of Theorem 4, meaning that* $\Delta = \mathrm{vec}(U)^{\otimes} := u^{\otimes l}$. *By the same logic of* (4.61), *we have that*

$$\begin{aligned}\mathbf{a}_1(\hat{\mathbf{w}}) &= \langle\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}})\rangle - (M^*)^{\otimes l}\rangle, \Delta\otimes\Delta\rangle \\ &= \underbrace{\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \hat{\mathbf{w}}\otimes\hat{\mathbf{w}}\otimes\Delta\otimes\Delta\rangle\rangle}_{\mathbf{b}_1} - \langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\otimes\Delta\otimes\Delta\rangle\rangle\end{aligned}$$

Since $\hat{\mathbf{w}}$ *is a* κ*-rank-1 tensor, by denoting* $\lambda_S\hat{x}^{\otimes l} := \mathbf{w}_\sigma$, *we represent*

$$\begin{aligned}\mathbf{b}_1 =& \langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \mathbf{w}_\sigma\otimes\mathbf{w}_\sigma\otimes\Delta\otimes\Delta\rangle\rangle + \\ & 2\underbrace{\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \mathbf{w}_\sigma\otimes\hat{\mathbf{w}}^\dagger\otimes\Delta\otimes\Delta\rangle\rangle}_{\mathbf{c}_1} + \\ & \underbrace{\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \hat{\mathbf{w}}^\dagger\otimes\hat{\mathbf{w}}^\dagger\otimes\Delta\otimes\Delta\rangle\rangle}_{\mathbf{c}_2}\end{aligned}$$

Hence,

$$\mathbf{a}_1(\hat{\mathbf{w}}) = \mathbf{a}_1(\mathbf{w}_\sigma) + 2\mathbf{c}_1 + \mathbf{c}_2$$

Now, we turn to $\mathbf{a}_2(\hat{\mathbf{w}})$. *Since the sensing matrices are assumed to be symmetric, by* (4.28), *we have*

$$\begin{aligned}\mathbf{a}_2(\hat{\mathbf{w}}) &= 4\langle\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\Delta)\rangle, \Delta\otimes\hat{\mathbf{w}}\rangle \\ &= 4\underbrace{\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \hat{\mathbf{w}}\otimes\Delta\otimes\hat{\mathbf{w}}\otimes\Delta\rangle\rangle}_{\mathbf{b}_2}\end{aligned}$$

again following the procedures in (4.61). *Given the decomposition of* $\hat{\mathbf{w}}$, *we decompose* $\mathbf{b}_2$ *similarly to* $\mathbf{b}_1$*:*

$$\begin{aligned}\mathbf{b}_2 =& \langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \mathbf{w}_\sigma\otimes\Delta\otimes\mathbf{w}_\sigma\otimes\Delta\rangle\rangle + \\ & \underbrace{\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \mathbf{w}_\sigma\otimes\Delta\otimes\hat{\mathbf{w}}^\dagger\otimes\Delta + \hat{\mathbf{w}}^\dagger\otimes\Delta\otimes\mathbf{w}_\sigma\otimes\Delta\rangle\rangle}_{\mathbf{c}_3} + \\ & \underbrace{\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, \langle\langle\mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l}\rangle, \hat{\mathbf{w}}^\dagger\otimes\Delta\otimes\hat{\mathbf{w}}^\dagger\otimes\Delta\rangle\rangle}_{\mathbf{c}_4}\end{aligned}$$

Combining everything together, we have

$$\begin{aligned}\nabla^2 h^l(\hat{\mathbf{w}})[\Delta, \Delta] &= \mathbf{a}_1(\mathbf{w}_\sigma) + 2\mathbf{c}_1 + \mathbf{c}_2 + \mathbf{a}_2(\mathbf{w}_\sigma) + 4\mathbf{c}_3 + 4\mathbf{c}_4 \\ &= \nabla^2 h^l(\mathbf{w}_\sigma)[\Delta, \Delta] + 2\mathbf{c}_1 + \mathbf{c}_2 + 4\mathbf{c}_3 + 4\mathbf{c}_4\end{aligned}$$

In addition, following the same procedures in (4.62),

$$2\mathbf{c}_1 + \mathbf{c}_2 + 4\mathbf{c}_3 + 4\mathbf{c}_4 \le (10\kappa + 5\kappa^2)\lambda_S^2 r^l \|\mathbf{A}^*\mathbf{A}\|_*^l$$

Now, since $\mathbf{w}_\sigma$ *is a lifted version of FOP for* (1.3) *(via Proposition 2),*

$$\nabla^2 h^l(\mathbf{w}_\sigma)[\Delta, \Delta] \leq -2G^l + \frac{2}{2^{l-1}} L_s^l \lambda_r(\hat{X}\hat{X}^\top)^l$$

where $\hat{X} = \mathrm{mat}(\hat{x})$ *and* $G := -\lambda_{min}(\nabla f(\hat{X}\hat{X}^\top)) \geq 0$*. Remember that the choice of* Δ *is identical. Therefore, a sufficient condition for* $\nabla^2 h^l(\hat{\mathbf{w}})[\Delta, \Delta] \leq 0$ *is that*

$$2G^l \geq \frac{2}{2^{l-1}} L_s^l \lambda_r(\hat{X}\hat{X}^\top)^l + (10\kappa + 5\kappa^2)\lambda_S^2 r^l \|\mathbf{A}^*\mathbf{A}\|_*^l$$

We can derive another sufficient condition to the above inequality, which is

$$G \geq 2^{1/l-1} L_s \lambda_r(\hat{X}\hat{X}^\top) + (5\kappa + 5\kappa^2/2)^{1/l} \lambda_S^{2/l} r \|\mathbf{A}^*\mathbf{A}\|_*$$

since $(a+b)^{1/l} \leq a^{1/l} + b^{1/l}$ *for* $a, b \geq 0$*. Following the steps of the proof of Theorem 4, we obtain that*

$$\|M^* - \hat{X}\hat{X}^\top\|_F^2 > 2^{1/l} \frac{L_s}{\alpha_s} \lambda_r(\hat{X}\hat{X}^\top)\, \mathrm{tr}(M^*) + \mathcal{O}(r\kappa^{1/l})$$

is sufficient. Note that $\|\mathbf{A}^*\mathbf{A}\|_*$ *can be rescaled to 1 easily. Following the same steps, we can set*

$$\beta = \frac{L_s\, \mathrm{tr}(M^*)\lambda_r(\hat{X}\hat{X}^\top)}{\alpha_s \|M^* - \hat{X}\hat{X}^\top\|_F^2 - \mathcal{O}(r\kappa^{1/l})}$$

and this leads to the desirable result.

4.C Custom Algorithms

Algorithm 1: CustomGD Algorithm

```
Input: learning_rate, n, r, l, prob_params, loss, g_thres, buffer, beta,
  gamma, eta_0
Initialize variables: A, b, escape_saddle, buffer_limit, buffer_step
Function init(starting_point, lr)
    if $lr \neq 0$ then
        learning_rate ← lr // Update learning rate if specified
    end
    return {'curr_iter' : 0,' t_noise' : 0,' curr_w' : starting_point}
Function update(gradients, opt_state)
    curr_iter ← opt_state['curr_iter'] + 1
    t_noise ← opt_state['t_noise']
    curr_w ← opt_state['curr_w']
    if $\|gradients\| < g\_thres$ and $curr\_iter > 100$ then
        if escape_saddle then
            t_noise ← curr_iter
            w_s ← find rank 1 component of curr_w using tensor PCA
            direction ← find the escape direction of w_s // According to
                Theorem 7
            this_eta ← eta_0
            while loss(curr_w + this_eta * direction) > loss(curr_w) + beta
              * this_eta * inner_product(gradients, direction) do
                this_eta ← this_eta * gamma // Update eta using gamma,
                    backtracking line search
            end
            updates ← this_eta * direction
            escape_saddle ← False
        end
        else
            buffer_step ← buffer_step + 1
            if buffer_step == buffer_limit then
                escape_saddle ← True
                buffer_step ← 0
            end
            updates ← -learning_rate * gradients
        end
    end
    else
        escape_saddle ← False
        updates ← -learning_rate * gradients
    end
    return updates, {'curr_iter' : curr_iter,' t_noise' : t_noise,' curr_w' :
      curr_w + updates}
```

Algorithm 2: Tensor PCA Algorithm

```
Input: tensor, lr, epochs, gradnorm_epsilon, lambd_v, key
Function tensor_PCA(tensor, lr, epochs, gradnorm_epsilon, lambd_v, key)
    Function loss(eigenval_eigenvec, tensor)
        lambd, v ← eigenval_eigenvec
        k ← len(tensor.shape)
        for each element in tensor.shape do
            tensor ← inner(tensor, v)
        end
        first_term ← square(lambd) * power(norm(v), 2*k)
        res ← first_term - 2*lambd*tensor
        return res
    s ← tensor.shape[0]
    if lambd_v is None then
        v ← random.normal(shape=(s,)) / sqrt(s)
        lambd ← 0.001 * random.normal()
    end
    else
        lambd, v ← lambd_v
    end
    loss, grads, lambd_v ← adam_optimize((loss, (lambd, v), tensor), lr,
      epochs, gradnorm_epsilon)
    lambd, v ← lambd_v
    sign ← sign(lambd)
    return sign * power(abs(lambd), 1 / len(tensor.shape)) * v
```

Chapter 5

Modified Loss

5.1 Background and Related Work

The optimization landscape of non-convex problems is notoriously complex to analyze in general due to the existence of an arbitrary number of spurious solutions (a spurious solution is a second-order critical point that is not a global minimum). As a result, if a numerical algorithm is not initialized close enough to a desirable solution, it may converge to one of those problematic spurious solutions. It may be acceptable (depending on the application) if the algorithm finds a critical point different from but close to the true solution, while converging to a point faraway implies the failure of the algorithm. In this chapter, we study this issue by focusing on (1.3), and provide an analysis of what can be done without the introduction of drastic over-parametrization like those elaborated in the previous two chapters.

The matrix sensing problem (1.3) is a canonical problem bearing many important applications, such as the matrix completion problem/netflix problem [15, 16], the compressed sensing problem [21], the training of quadratic neural networks [48], and an array of localization/estimation problems [104, 36, 75, 8, 74, 23]. As a result, a better understanding of (1.3) not only helps with the above applications, but also paves the way for the analysis of a broader range of non-convex problems. This is due in part to the fact that any polynomial optimization can be converted into a series of matrix sensing problems under benign assumptions [61].

As previous discussed, the major drawback of (1.3) is that it may have an arbitrary number of spurious solutions, which cause ubiquitous local search algorithms to potentially end up with unwanted solution. Therefore, there has been an extensive investigation of the non-convex optimization landscape of (1.3), and the centerpiece notion is the restricted isometry property (RIP) given in Definition 1. Intuitively speaking, a smaller RIP constant means that the problem is easier to solve. For instance, if $\delta_{2r} = 0$, then $\mathcal{A}(\cdot)$ becomes the identity operator with $b = \text{vec}(M^*)$, which makes the problem trivial to solve for M^*.

5.1.1 Related Works

In this section, we delve into the optimization landscape of the non-convex problem delineated in equation (1.3), highlighting the pivotal role played by the Restricted Isometry Property (RIP). Despite the extensive literature addressing this topic, we argue for the necessity of further investigation. We structure our discussion around two scenarios: the RIP constant being less than 1/2 and greater than 1/2, to provide a clearer and more organized review.

The significance of the RIP constant in influencing the optimization landscape was first brought to light through the application of convex semidefinite programming (SDP) relaxations in matrix sensing challenges, as evidenced by seminal works [72, 16]. These studies demonstrated that a condition of $\delta_{5r} \leq 1/10$ guarantees the exact recovery of M^* via SDP relaxation. Further exploration by [5] on the factorized version of the problem, represented in (1.3), revealed that a tighter RIP constraint, $\delta_{2r} \leq 1/5$, ensures that all second-order critical points (SOPs) are indeed the true solutions. This was corroborated by subsequent research [105, 46], which extended the sufficiency of this RIP constraint to arbitrary objective functions for global recovery of M^*. A novel approach by [101], employing a "certification of in-existence" technique, established $\delta_{2r} = 1/2$ as a critical threshold. This finding indicates that below this threshold, (1.3) is devoid of misleading solutions, whereas exceeding it introduces potential counter-examples with SOPs that diverge from the true solutions. These results are achieved in the exact-parametrized space, namely when $r_{\text{search}} = r$.

The complexity of the problem escalates significantly as the RIP constant surpasses the 1/2 threshold. [101]'s investigation into cases where $\delta_{2r_{\text{search}}} \geq 1/2$ demonstrated that spurious solutions can be avoided locally around M^*, with the feasibility dependent on the RIP constant and the dimensions of M^*. An alternative strategy, over-parametrization, involves setting $r_{\text{search}} > r$, and was proven by [97] to ensure that every SOP $\hat{X}$ satisfies $\hat{X}\hat{X}^\top = M^*$. This is contingent upon choosing r such that it satisfies a specific relation with r and the RIP constant. Further insights into addressing the challenges presented by high RIP constants in under-sampled scenarios are provided through innovative methods such as SDP relaxation and tensor space lifting, detailed in Chapters 3 and 4 respectively.

Despite these advancements, the optimization landscape of (1.3) under a high RIP constant remains a complex issue. The proposed solutions either necessitate significantly increased algorithmic complexity (via methods such as over-parametrization, SDP relaxation, or tensor optimization) or require initialization near the true solution, M^*. This observation leads us to question: *Can we achieve global guarantees for* (1.3) *with* $\delta \geq 1/2$ *without substantially escalating the computational complexity?* This chapter seeks to provide a partial affirmative response through the introduction of high-order losses.

5.2 Global Landscape of Matrix Sensing

As discussed previously, the optimization landscape of (1.3) is benign (in the sense of having no spurious solutions) if $\delta_{2r} < 1/2$ and benign in a region close to M^* if $\delta_{2r} \geq 1/2$. In this section, we study the landscape far away from M^* in the problematic case $\delta_{2r} \geq 1/2$. To do so, we focus on the first-order critical points, and study the eigenvalues of the Hessian at these points because if they exhibit negative eigenvalues, it means that these first-order critical points are strict saddles, possessing escape directions. Please recall Lemma 4 for the definition and form of the FOPs and SOPs of this problem. Focusing on (2.6), it is apparent that for a first-order critical point $\hat{X}$ satisfying $\nabla h(\hat{X}) = 0$, the Hessian $\nabla^2 h(\hat{X})[U, U]$ can be broken down into the summation of two terms:

$$\begin{aligned} T_1 &:= \sum_{i=1}^{m} \langle A_i, U\hat{X}^\top + \hat{X}U^\top \rangle^2 = \|\mathcal{A}(U\hat{X}^\top + \hat{X}U^\top)\|_2^2, \\ T_2 &:= \sum_{i=1}^{m} \langle A_i, \hat{X}\hat{X}^\top - M^* \rangle \langle A_i, 2UU^\top \rangle \\ &= 2\langle \nabla f(\hat{X}\hat{X}^\top), UU^\top \rangle \end{aligned}$$

Assuming that the problem (1.3) satisfies the RIP condition with some constant δ_p for $p \geq 2r$, one can write

$$T_1 \leq (1 + \delta_p)\|U\hat{X}^\top + \hat{X}U^\top\|_F^2,$$

which means that T_1 can be upper-bounded naturally since U can be assumed to have a unit scale without loss of generality. Therefore, if we can somehow show that there exists $U \in \mathbb{R}^{n \times r_{\text{search}}}$ to make T_2 negative with a sufficiently large magnitude, then $\hat{X}$ becomes a saddle point. Combining the RIP condition and mean value theorem, we know that

$$\begin{aligned} f(M^*) \geq & f(\hat{X}\hat{X}^\top) + \langle \nabla f(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top \rangle \\ & + \frac{1 - \delta_p}{2}\|\hat{X}\hat{X}^\top - M^*\|_F^2 \end{aligned}$$

Given $\nabla h(\hat{X}) = 0$ and the expression in (2.5), we obtain that

$$\langle \nabla f(\hat{X}\hat{X}^\top), M^* \rangle \leq -\frac{1 - \delta_p}{2}\|\hat{X}\hat{X}^\top - M^*\|_F^2$$

since $f(\hat{X}\hat{X}^\top) \geq f(M^*)$ by definition. This implies that there exist directions that make T_2 have large negative values when $\hat{X}\hat{X}^\top$ is far away from M^*. Expanding on this simple observation, we formally establish Theorem 8, serving as the cornerstone of all results in this paper. A detailed proof can be found in the Appendix.

Theorem 8 *Assume that* (1.3) *satisfies the RIP*$_{r_{search}+r}$ *property with constant* $\delta \in [0,1)$. *Given a first-order critical point* $\hat{X} \in \mathbb{R}^{n\times r_{search}}$ *of* (1.3), *if it satisfies the inequality*

$$\|\hat{X}\hat{X}^\top - M^*\|_F^2 > 2\frac{1+\delta}{1-\delta}\operatorname{tr}(M^*)\sigma_r(\hat{X})^2, \tag{5.1}$$

then $\hat{X}$ *is not a second-order critical point and is a strict saddle point with* $\nabla^2 h(\hat{X})$ *having a strictly negative eigenvalue not larger than*

$$2(1+\delta)\sigma_r(\hat{X})^2 - \frac{\|\hat{X}\hat{X}^\top - M^*\|_F^2(1-\delta)}{\operatorname{tr}(M^*)} \tag{5.2}$$

Theorem 8 states that if a first-order critical point is far from the ground truth, it cannot be a spurious solution, and always exhibits an escape direction with its magnitude proportional to the squared distance between $\hat{X}\hat{X}^\top$ and M^*. This further elucidates the fact that even if (1.3) is poorly initialized, it is possible to converge to a vicinity of M^* with saddle-escaping algorithms, which we will numerically illustrate in Section 5.4.

5.3 Higher-Order Loss Functions

Although Theorem 8 proves that critical points far away from the ground truth are strict saddle points, the time needed to escape such points depends on the local curvature of the function [24, 33]. Therefore, it is essential to understand whether the curvatures at saddle points could be enhanced to reshape the landscape favorably. In this section, we provide an affirmative answer to this question by using a modified loss function.

Our main goal in matrix sensing is to recover the ground truth matrix M^* via m measurements, and we minimize a mismatch error in (1.3) to achieve this goal. An l_2 loss function is used in (1.3) due to its smooth and nonnegative properties, which is the most common objective function in the machine learning literature. However, in this work, we introduce a high-order loss function as penalization, namely an l_p loss function with $p > 2$, and show that this will reshape the landscape of the optimization problem. To be concrete, we propose to optimize over this modified problem:

$$\min_{X\in\mathbb{R}^{n\times r_{\text{search}}}} h_\lambda^l(X) := h(X) + \lambda h^l(X) \tag{5.3}$$

where

$$h^l(X) := \frac{1}{l}\|\mathcal{A}(XX^\top) - b\|_l^l \tag{5.4a}$$

$$f^l(M) := \frac{1}{l}\|\mathcal{A}(M) - b\|_l^l \tag{5.4b}$$

where $l \geq 2$ is an even natural number to ensure the non-negativity of the loss function and $\lambda > 0$ is a penalty coefficient. The intuition behind using a high-order objective can be easily demonstrated via the scalar example:

$$\min_{x\in\mathbb{R}} g(x) := \frac{1}{l}(x^2 - a)^l \tag{5.5}$$

for some constant $a \in \mathbb{R}$ and an even number $l \geq 2$. This problem is a scalar analogy of $h^l(X)$ with $\mathcal{A}(\cdot)$ being the identity operator. In this example, the derivatives are

$$g'(x) = 2x(x^2 - a)^{l-1},$$
$$g''(x) = 2(x^2 - a)^{l-2}\left[(l-1)2x^2 + (x^2 - a)\right]$$

It can be observed that as l increases, the first- and second-order derivatives will be amplified, provided that $(x^2 - a)$ is larger than one (i.e., our point is reasonably distant from the ground truth a). However, there is an issue with optimizing $g(x)$ directly, and we need to use $h^l_\lambda(X)$ instead of $h^l(X)$. If we directly minimize $h^l(X)$ with $l > 2$, the Hessian at any point X with the property $XX^T = M^*$ becomes zero, which makes the convergence extremely slow as approaching the ground truth (with a sub-linear rate). This is because the local convergence rate of descent numerical algorithms depends on the condition number of the Hessians around the solution [86]. Conversely, when using the original objective (1.3), we see from (2.6) that even if $XX^\top$ is close to M^*, the Hessian is positive semidefinite, and therefore adding $h^l(X)$ to the objective of (1.3) will not change the sign of the Hessian around the solution.

Secondly, if l is large and $\|XX^\top - M^*\|_F$ is less than one, the term $\langle A_i, XX^\top - M^*\rangle^{l-1}$ appearing in the gradient of $h^l(X)$ (see Lemma 20 in Appendix) is very small due to its exponentiation nature. This means that minimizing $h^l(X)$ alone will suffer from the vanishing issue and slow growth rate in a local region around M^*.

Due to the above reasons, we mix $h^l(X)$ with the original objective in (1.3) and use the parameter λ to control the effect of the penalty term, in an effort to balance local rate of convergence to M^* and prominent eigenvalues of the Hessian at points far away from M^*. By using (5.3), we can arrive at a similar result to Theorem 8

Theorem 9 *Assume that the operator $\mathcal{A}(\cdot)$ satisfies the RIP$_{r_{search}+r}$ property with constant $\delta \in [0, 1)$. Consider the high-order optimization problem (5.3) such that $l \geq 2$ is even. Given a first-order critical point $\hat{X} \in \mathbb{R}^{n\times r_{search}}$ of (5.3), if*

$$D^2 \geq \operatorname{tr}(M^*)\sigma_r^2(\hat{X})\frac{(1+\delta) + \lambda(l-1)(1+\delta)^{l/2}D^{l-2}}{(1-\delta)/2 + \lambda C(l)(1-\delta)^{l/2}D^{l-2}}, \tag{5.6}$$

then $\hat{X}$ is a strict saddle point with $\nabla^2 h(\hat{X})$ having a strictly negative eigenvalue not *larger than*

$$\begin{aligned}&\left[2(1+\delta)\sigma_r(\hat{X})^2 - \frac{D^2(1-\delta)}{\operatorname{tr}(M^*)}\right]+\\ \lambda D^{l-2}&\left[2(1+\delta)^{l/2}(l-1)\sigma_r(\hat{X})^2 - 2\frac{(1-\delta)^{l/2}C(l)D^2}{\operatorname{tr}(M^*)}\right]\end{aligned} \tag{5.7}$$

where

$$\begin{aligned}D &:= \|\hat{X}\hat{X}^\top - M^*\|_F,\\ C(l) &:= m^{(2-l)/2}\left(\frac{2^l-1}{l}-1\right)\end{aligned} \tag{5.8}$$

Theorem 9 serves as a direct generalization of Theorem 8, as it recovers the statements of Theorem 8 when l is set to 2 or λ is set to 0. By comparing (5.7) to (5.2), the bound on the smallest eigenvalue of the Hessian has an additional term that is amplified by $\|\hat{X}\hat{X}^\top - M^*\|_F^{l-2}$. As a result, $\hat{X}$ has a more pronounced escape direction in (5.3) compared to (1.3) when $\|\hat{X}\hat{X}^\top - M^*\|_F$ is large. Concerning the tightness of the bounds in Theorem 8 and Theorem 9, they depend on three factors: the RIP constant, the Frobenius norm of M^*, and the smallest non-zero singular value of $\hat{X}$. While the RIP constant indicates problem difficulty and is immutable, knowing that $\hat{X}$ has a minimal singular value (which is computable), suggests that the bounds can be very tight. This implies that if $\hat{X}$ is a second-order point, it will be very close to M^*. Recent studies [76, 34] have shown that a small random initialization can lead to small r^{th} eigenvalues during the optimization process, resulting in tight bounds for our new Theorems.

Theorem 9 contrasts well with our approach introduced in Chapter 4. Theorem 4 states that by lifting the search space to the regime of tensors, a higher degree of parametrization can amplify the negative curvature of Hessian. In contrast, Theorem 9 offers similar benefits by using a more complex objective function. This means that without resorting to massive over-parametrization, similar results can be achieved via using a more complex loss function. Having said that, the technique presented in [56] can amplify the negative curvature of those points X that satisfy

$$\|XX^\top - M^*\|_F^2 \geq \frac{1+\delta}{1-\delta}\operatorname{tr}(M^*)\sigma_r^2(\hat{X})$$

where in comparison to (5.6) the multiplicative factor to $\operatorname{tr}(M^*)\sigma_r^2(\hat{X})$ becomes

$$\frac{(1+\delta)+\lambda(l-1)(1+\delta)^{l/2}D^{l-2}}{(1-\delta)/2+\lambda C(l)(1-\delta)^{l/2}D^{l-2}},$$

which is on the order of magnitude of

$$\mathcal{O}\left(l\left(\frac{\sqrt{m}}{2}\right)^l\left(\frac{1+\delta}{1-\delta}\right)^{l/2}\right),$$

n	λ	$\lambda_{\min}(\nabla^2 h^l(\hat{X}))$	$\lambda_{\max}(\nabla^2 h^l(\hat{X}))$	$\lambda_{\min}(\nabla^2 f^l(X^*))$	$\lambda_{\max}(\nabla^2 f^l(X^*))$
5	0	0.429	3.898	0.54	4.72
5	0.5	0.421	4.106	0.54	4.72
5	5	0.385	9.117	0.54	4.72
5	50	0.354	69.816	0.54	4.72
7	0	0.516	3.642	0.72	5.08
7	0.5	0.502	4.122	0.72	5.08
7	5	0.456	10.006	0.72	5.08
7	50	0.433	75.786	0.72	5.08
9	0	0.609	3.930	0.90	5.44
9	0.5	0.601	4.315	0.90	5.44
9	5	0.557	10.915	0.90	5.44
9	50	0.528	84.002	0.90	5.44

Table 5.1: The smallest eigenvalue of the Hessian at a spurious local minimum $\hat{X}$ and ground truth X^*, with $\epsilon = 0.3$ and additional high-order loss function $l = 4$ (note that $\hat{X}$ is not too far from X^* since Theorem 8 shows that there are no such spurious solutions). The problem satisfies the $RIP_{2r_{\text{search}}}$-property with $\delta = \frac{1-\varepsilon}{1+\varepsilon} = 0.538 > 1/2$, and hence has spurious local minima.

making the region for which this amplification can be observed smaller if l is large. This means that by utilizing a high-order loss, we can recover some of the desirable properties of an over-parametrized technique, but a gap still exists due to the smaller parametrization. Combining a high-order loss function and a modest level of parametrization is left as future work.

5.4 Numerical Experiments

This section serves to provide numerical validation for the theoretical findings presented in this paper. We will begin by investigating the behavior of the Hessian matrix when utilizing high-order loss function. Subsequently, we will showcase the remarkable acceleration in escaping saddle points achieved by employing Perturbed Gradient Descent in conjunction with high-order loss functions compared to the standard optimization problem (1.3). Lastly, we will provide a comparative illustration of the landscape both with and without the incorporation of high-order loss functions.

We first focus on a benchmark matrix sensing problem with the operator $\mathcal{A}$ defined as

$$\mathcal{A}_\epsilon(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij}, & \text{if } (i,j) \in \Omega \\ \epsilon \mathbf{M}_{ij}, & \text{otherwise} \end{cases}, \tag{5.9}$$

where $\Omega = \{(i,i), (i,2k), (2k,i) \mid \forall i \in [n], k \in [\lfloor n/2 \rfloor]\}$, $0 < \epsilon < 1$. [88] has proved that while satisfying RIP property with $\delta_{2r_{\text{search}}} = (1-\epsilon)/(1+\epsilon)$, this problem has $\mathcal{O}\left(2^{\lceil n/2 \rceil} - 2\right)$ spurious local minima. In order to analyze the influence of high-order loss functions on the optimization landscape, we conduct an analysis of the spurious local minima and of the ground truth matrix for both the vanilla problem (1.3) and the altered problem (5.3) with $l = 4$. We consider a spurious local minimum $\hat{X}$ (note that such points cannot be too far away from M^* due to Theorem 8). The findings of this study are presented in Table 5.1, while the ratio between the largest and smallest eigenvalues at the spurious local minimum $\hat{X}$ is plotted in Figure 5.1.

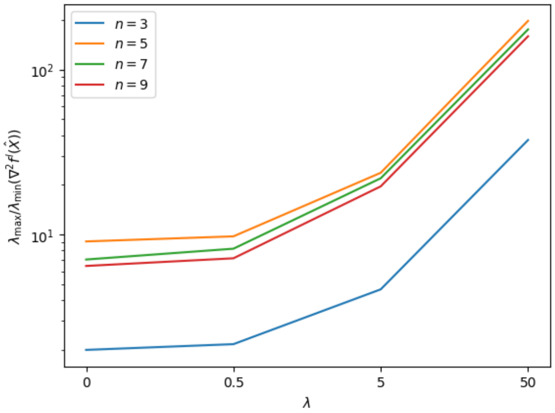
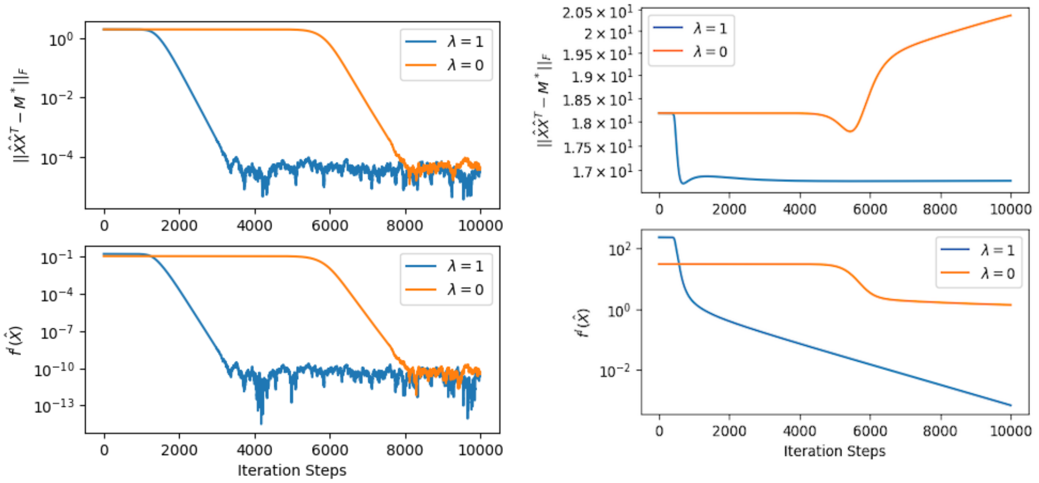


Figure 5.1: The ratio between the largest and smallest eigenvalue of Hessian at the spurious local minimum $\lambda_{\max}/\lambda_{\min}(\nabla^2 h^l(\hat{X}))$ with respect to λ under different size n.

Table 5.1 shows that as the intensity of high-order loss function increases via λ, the behavior of the Hessian eigenvalues exhibits distinct characteristics across different points in the optimization landscape. Specifically, the smallest and the largest eigenvalues of the Hessian at the ground truth matrix remain constant. In contrast, the smallest eigenvalue of the Hessian at the spurious local minimum, which is initially positive, decreases as λ increases. This decreasing trend facilitates the differentiation of spurious local minima from the global minimum, as they become less favorable. Simultaneously, the largest eigenvalue of the Hessian matrix at the spurious local minimum increases at a significantly faster rate. This suggests that the incorporation of high-order loss functions amplifies the magnitude of the eigenvalues in the Hessian matrix at spurious local minima, increasing the ratio between the largest and smallest eigenvalues, while having no impact on those at the ground truth.

Following that, we will present the acceleration in effectively navigating away from saddle points. For randomly generated zero-mean Gaussian sensing matrices with i.i.d. entries, we apply small initialization and perturbed gradient descent which adds small Gaussian noise when the gradient is close to zero. In Figure 5.2, we compare the evolution of the distance from the ground truth matrix $\|\hat{X}\hat{X}^T - M^*\|_F$

and the value of the objective function $h^l(\hat{X})$. Although Figure 5.2 demonstrates the behavior for a single problem, we observed the same phenomenon for many different trials. By incorporating a high-order loss function (specifically, with $l = 4$), the optimization process exhibits enhanced convergence compared to the standard vanilla problem. This accelerated convergence can be attributed to the presence of a substantial negative eigenvalue of the Hessian matrix, which effectively facilitates the algorithm's escape from regions proximate to spurious local minimum.



(a) $\lambda = 0$ converges to ground truth

(b) $\lambda = 0$ converges to a spurious solution around the ground truth

Figure 5.2: The evolution of the objective function and the error between the obtained solution $\hat{X}\hat{X}^T$ and the ground truth M^* during the iterations of the perturbed gradient descent method, with a constant step-size. In both cases, high-order loss functions accelerate the convergence.

Finally, we explore the optimization landscape in terms of the distance from the ground truth matrix and the intensity of high-order loss functions. We explore both random Gaussian sensing matrices with size $m = 20, n = 20$, and the problem (5.9) with the parameters $n = 21$ and $l = 4$. The result is plotted in Figure 5.3, where the x-axis and y-axis are two orthogonal directions from the critical point to the ground truth. By looking horizontally across the figure, we can observe that increasing the parameter λ leads to the amplification of the least negative eigenvalue of the Hessian matrix at saddle points. As λ increases, the least eigenvalue of the Hessian at this saddle point, which is initially negative, decreases further. This reduction in the magnitude of the negative eigenvalue makes it easier to escape from this saddle point during optimization.

This example could also directly corroborate Theorem 9 (Thereby Theorem 8). For instance, when $\lambda = 0.5$, the minimum eigenvalue of Hessian matrix at the first-order critical point $\hat{X}$ is $\lambda_{\min}(\nabla^2 h^l(\hat{X})) = -3.201$, which is smaller than the

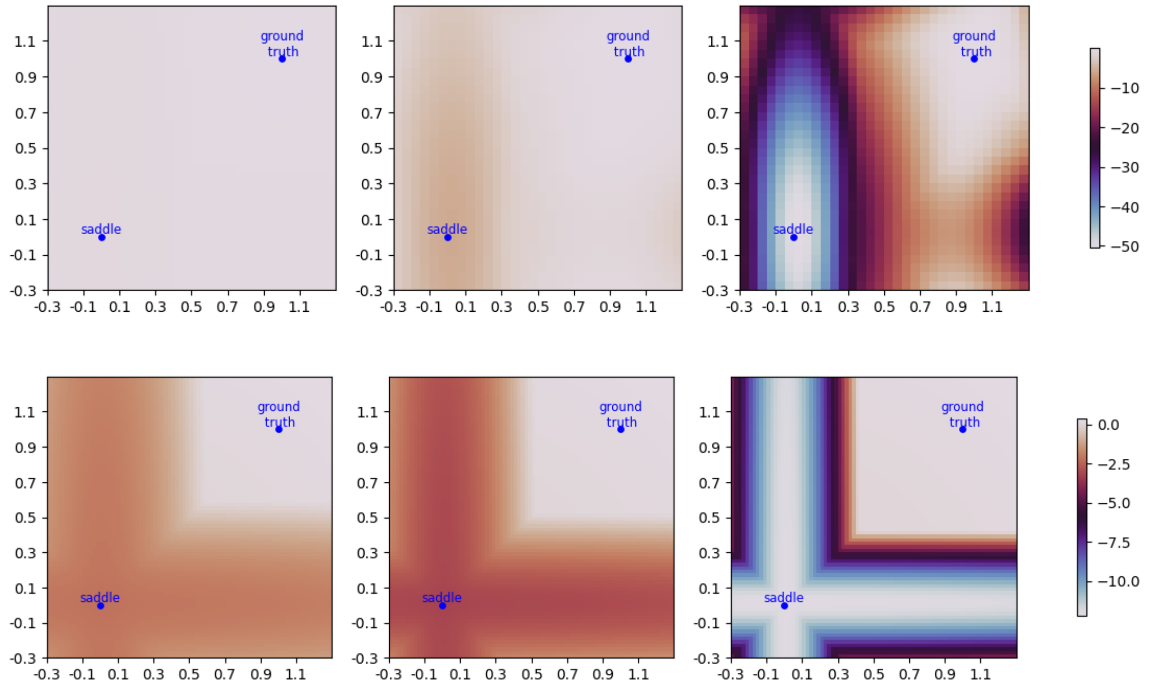


Figure 5.3: The value of the minimum eigenvalue of the Hessian around saddle points: The first row is for randomly generated Gaussian matrix with $m = 20, n = 20$, and the second row is for problem (5.9) with $n = 21, \epsilon = 0.1$. $\lambda = 0$ (left column), $\lambda = 0.5$ (middle column), $\lambda = 5$ (right column), with x-axis and y-axis as two orthogonal directions from the critical point to the ground truth.

eigenvalue-bound -2.274; the distance from the ground truth matrix is $D := \|\hat{X}\hat{X}^\top - M^*\|_F = 11.0$, larger than the distance bound in (5.6), validating Theorem 9.

5.5 Summary

This chapter theoretically establishes favorable geometric properties in those parts of the space far from the globally optimal solution for the non-convex matrix sensing problem. We introduce the notion of high-order loss functions and show that such losses reshape the optimization landscape and accelerate escaping saddle points. Our experiments demonstrate that high-order penalties decrease minimum Hessian eigenvalues at spurious points while intensifying ratios. Secondly, perturbed gradient descent exhibits accelerated saddle escape with the incorporation of high-order losses. Collectively, our theoretical and empirical results show that using a modified loss function could make non-convex functions easier to deal with and achieve some of the desirable properties of a lifted formulation without enlarging the search space of the problem exponentially.

5.A Missing Details of Chapter 5

Lemma 20 *Given the problem* (5.3)*, its gradient and Hessian are given as*

$$\langle \nabla h^l(X), U\rangle = \sum_{i=1}^{m} \langle A_i, XX^\top - M^*\rangle^{l-1} \langle A_i, UX^\top + XU^\top\rangle \quad \forall U \in \mathbb{R}^{n\times r_{search}}, \tag{5.10}$$

and

$$\begin{aligned} &\nabla^2 h^l(X)[U,U] = \sum_{i=1}^{m} \langle A_i, XX^\top - M^*\rangle^{l-2} \\ &[(l-1)\langle A_i, UX^\top + XU^\top\rangle^2 + \langle A_i, XX^\top - M^*\rangle\langle A_i, 2UU^\top\rangle] \quad \forall U \in \mathbb{R}^{n\times r_{search}} \end{aligned} \tag{5.11}$$

Lemma 21 *Given the problem* (5.4b)*, its gradient, Hessian and high-order derivatives are equal to*

$$\nabla f^l(M) = \sum_{i=1}^{m} \langle A_i, M - M^*\rangle^{l-1} A_i, \tag{5.12a}$$

$$\nabla^2 f^l(M)[N,N] = (l-1)\sum_{i=1}^{m} \langle A_i, M - M^*\rangle^{l-2} \langle A_i, N\rangle^2 \quad \forall N \in \mathbb{R}^{n\times n}, \tag{5.12b}$$

$$\nabla^p f^l(M)[\underbrace{N, \dots, N}_{p\ times}] = \frac{(l-1)!}{(l-p)!}\sum_{i=1}^{m} \langle A_i, M - M^*\rangle^{l-p} \langle A_i, N\rangle^p \quad \forall N \in \mathbb{R}^{n\times n} \tag{5.12c}$$

Proof to both lemmas are simply multivariate calculus, thus obviated here for simplicity.

Proof 23 (Proof of Theorem 8) *Via the definition of RIP, we have that*

$$f(M^*) \geq f(\hat{X}\hat{X}^\top) + \langle \nabla f(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top\rangle + \frac{1-\delta}{2}\|\hat{X}\hat{X}^\top - M^*\|_F^2$$

Given that $\hat{X}$ is a first-order critical point, it means that it must satisfy first-order optimality conditions, which means

$$f(\hat{X}\hat{X}^\top)\hat{X} = 0,$$

leading to

$$\langle \nabla f(\hat{X}\hat{X}^\top), M^*\rangle \leq -\frac{1-\delta}{2}\|\hat{X}\hat{X}^\top - M^*\|_F^2$$

since $f(\hat{X}\hat{X}^\top) - f(M^*) \geq 0$ *via the construction of the objective function. Furthermore, as it is without loss of generality to assume the gradient of* $h(M)$ *to be symmetric [94], and the fact that* M^* *is positive-semidefinite, we have that*

$$\langle \nabla f(\hat{X}\hat{X}^\top), M^* \rangle \geq \lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \operatorname{tr}(M^*)$$

This leads to the fact that

$$\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \leq -\frac{(1-\delta)\|\hat{X}\hat{X}^\top - M^*\|_F^2}{2\operatorname{tr}(M^*)} \leq 0 \tag{5.13}$$

Given the optimality conditions, we know that for $\hat{X}$ *to be a strict saddle, we must prove that there exists a direction* $\Delta \in \mathbb{R}^{n \times r_{search}}$ *such that*

$$2\langle \nabla f(\hat{X}\hat{X}^\top), \Delta\Delta^\top \rangle + \underbrace{\left[\nabla^2 f(\hat{X}\hat{X}^\top)\right](\hat{X}\Delta^\top + \Delta\hat{X}^\top, \hat{X}\Delta^\top + \Delta\hat{X}^\top)}_{P(\Delta)} < 0 \tag{5.14}$$

If we choose

$$\Delta = uq^\top, \quad \|u\|_2, \|q\|_2 = 1, \quad \|\hat{X}q\|_2 = \sigma_r(\hat{X}),\ u^\top \nabla f(\hat{X}\hat{X}^\top)u = \lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)),$$

then it follows from the RIP condition that

$$\begin{aligned} P(\Delta) &\leq (1+\delta)\|\hat{X}\Delta^\top + \Delta\hat{X}^\top\|_F^2 \\ &= (1+\delta)\|u(\hat{X}q)^\top + (\hat{X}q)u^\top\|_F^2 \\ &= 2(1+\delta)\|\hat{X}q\|_F^2 + 2(1+\delta)\left(q^\top(\hat{X}^\top u)\right)^2 \\ &= 2(1+\delta)\sigma_r(\hat{X})^2 \end{aligned}$$

because of $\hat{X}^\top u = 0$ *due to the first-order optimality condition. Therefore,*

$$\begin{aligned} 2\langle \nabla f(\hat{X}\hat{X}^\top), \Delta\Delta^\top \rangle + P(\Delta) &\leq 2\langle f(\hat{X}\hat{X}^\top), \Delta\Delta^\top \rangle + 2(1+\delta)\sigma_r(\hat{X})^2 \\ &= 2\langle \nabla f(\hat{X}\hat{X}^\top), uu^\top \rangle + 2(1+\delta)\sigma_r(\hat{X})^2 \\ &= 2u^\top \nabla f(\hat{X}\hat{X}^\top)u + 2(1+\delta)\sigma_r(\hat{X})^2 \\ &= 2\left(\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) + (1+\delta)\sigma_r(\hat{X})^2\right) \\ &\leq 2(1+\delta)\sigma_r(\hat{X})^2 - \frac{(1-\delta)\|\hat{X}\hat{X}^\top - M^*\|_F^2}{\operatorname{tr}(M^*)} \end{aligned}$$

Therefore, in order to make (5.14) *hold, we simply need*

$$\|\hat{X}\hat{X}^\top - M^*\|_F^2 > 2\frac{1+\delta}{1-\delta}\operatorname{tr}(M^*)\sigma_r(\hat{X})^2$$

which concludes the proof.

Before proceeding to the main proof, we first present a technical lemma:

Lemma 22 *Given a vector $x \in \mathbb{R}^n$, we have that*

$$\|x\|_p \leq n^{1/p-1/q}\|x\|_q \quad \forall\ q \geq p \tag{5.15}$$

Proof 24 (Proof of Lemma 22) *By applying Holder's inequality, we obtain that*

$$\sum_{i=1}^n |x_i|^p = \sum_{i=1}^n |x_i|^p \cdot 1 \leq \left(\sum_{i=1}^n (|x_i|^p)^{\frac{q}{p}}\right)^{\frac{p}{q}} \left(\sum_{i=1}^n 1^{\frac{q}{q-p}}\right)^{1-\frac{p}{q}} = \left(\sum_{i=1}^n |x_i|^q\right)^{\frac{p}{q}} n^{1-\frac{p}{q}}$$

Then

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \leq \left(\left(\sum_{i=1}^n |x_i|^q\right)^{\frac{p}{q}} n^{1-\frac{p}{q}}\right)^{1/p} = \left(\sum_{i=1}^n |x_i|^q\right)^{\frac{1}{q}} n^{\frac{1}{p}-\frac{1}{q}} = n^{1/p-1/q}\|x\|_q$$

Proof 25 (Proof to Theorem 9) *First, we define*

$$f_\lambda^l(M) = f(M) + \lambda f^l(M)$$

Then, focusing on $h^l(\cdot)$, using Taylor's theorem with remainder, we get

$$\begin{aligned} f^l(M^*) =& f^l(\hat{X}\hat{X}^\top) + \langle \nabla f^l(\hat{X}\hat{X}^\top), \Delta\rangle + \frac{1}{2!}\nabla^2 f^l(\hat{X}\hat{X}^\top)\,[\Delta,\Delta] + \\ & \frac{1}{3!}\nabla^3 f^l(\hat{X}\hat{X}^\top)\,[\Delta,\Delta,\Delta] + \dots + \frac{1}{l!}\nabla^l f^l(\tilde{M})\,[\Delta,\dots,\Delta] \end{aligned} \tag{5.16}$$

where $\tilde{M}$ is a convex combination of M^* and $\hat{X}\hat{X}^\top$ and

$$\Delta = M^* - \hat{X}\hat{X}^\top.$$

Using Lemma 21, we know that

$$\begin{aligned} &\frac{1}{p!}\nabla^p f^l(\hat{X}\hat{X}^\top)\,[\Delta,\dots,\Delta] = \frac{(l-1)!}{(l-p)!p!}\sum_{i=1}^m \langle A_i,\Delta\rangle^l \quad \forall p \in [2, l-1], \\ &\frac{1}{l!}\nabla^l f^l(M)\,[\Delta,\dots,\Delta] = \frac{(l-1)!}{(l-l)!l!}\sum_{i=1}^m \langle A_i,\Delta\rangle^l = \frac{1}{l}\sum_{i=1}^m \langle A_i,\Delta\rangle^l \quad \forall M \in \mathbb{R}^{n\times n} \end{aligned}$$

Hence, it is possible to rewrite

$$\begin{aligned} f^l(M^*) &= f^l(\hat{X}\hat{X}^\top) + \langle \nabla f^l(\hat{X}\hat{X}^\top), \Delta\rangle + \sum_{p=2}^l \frac{(l-1)!}{(l-p)!p!}\sum_{i=1}^m \langle A_i,\Delta\rangle^l \\ &= f^l(\hat{X}\hat{X}^\top) + \langle \nabla f^l(\hat{X}\hat{X}^\top), \Delta\rangle + \sum_{i=1}^m \langle A_i,\Delta\rangle^l \sum_{p=2}^l \frac{(l-1)!}{(l-p)!p!} \\ &= f^l(\hat{X}\hat{X}^\top) + \langle \nabla f^l(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top\rangle + (\frac{2^l-1}{l} - 1)\|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_l^l \end{aligned} \tag{5.17}$$

By Lemma 22, if $l > 2$, it holds that

$$\|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_l^l \geq \|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_2^l m^{(2-l)/2}$$

Furthermore, combining this with the RIP property gives rise to

$$\|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_l^l \geq (1-\delta)^{l/2}\|\hat{X}\hat{X}^\top - M^*\|_F^l m^{(2-l)/2} \tag{5.18}$$

Therefore we know that

$$f^l(M^*) \geq f^l(\hat{X}\hat{X}^\top) + \langle \nabla f^l(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top\rangle + (1-\delta)^{l/2}C(l)\|\hat{X}\hat{X}^\top - M^*\|_F^l \tag{5.19}$$

if we summarize constant relevant to l as $C(l)$. Thus, using the above inequality twice, once with $l = 1$ and another with general l, we get

$$f_\lambda^l(M^*) = f(M^*) + \lambda f^l(M^*) \geq f_\lambda^l(\hat{X}\hat{X}^\top) + \langle \nabla f_\lambda^l(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top\rangle + L \tag{5.20}$$

in which

$$L := \frac{1-\delta}{2}\|M^* - \hat{X}\hat{X}^\top\|_F^2 + \lambda(1-\delta)^{l/2}C(l)\|M^* - \hat{X}\hat{X}^\top\|_F^l$$

If $\hat{X}$ is a first-order critical point, a repeated application of (5.10) *yields that*

$$\nabla f_\lambda^l(\hat{X}\hat{X}^\top)\hat{X} = 0 \implies \langle \nabla f_\lambda^l(\hat{X}\hat{X}^\top), \hat{X}\hat{X}^\top\rangle = 0$$

which after rearrangement leads to

$$\begin{aligned}\langle \nabla f_\lambda^l(\hat{X}\hat{X}^\top), M^*\rangle &\leq \left[f_\lambda^l(M^*) - f_\lambda^l(\hat{X}\hat{X}^\top)\right] - L \\ &\leq -L\end{aligned} \tag{5.21}$$

where the second inequality follows from the fact that M^* is the global minimizer of (5.3)*. Since the sensing matrices can be assumed to be symmetric without loss of generality [94], $\nabla f^l(\hat{X}\hat{X}^\top)$ can be assumed to be symmetric according to* (5.12a)*. This means that*

$$\langle \nabla f_\lambda^l(\hat{X}\hat{X}^\top), M^*\rangle \geq \operatorname{tr}(M^*)\lambda_{\min}(\nabla f_\lambda^l(\hat{X}\hat{X}^\top))$$

which further leads to

$$\lambda_{\min}(\nabla f_\lambda^l(\hat{X}\hat{X}^\top)) \leq -\frac{L}{\operatorname{tr}(M^*)} \tag{5.22}$$

Now, we turn to the Hessian of $f_\lambda^l(\cdot)$, which given (20) *is*

$$\begin{aligned}\nabla^2 h_\lambda^l(\hat{X})[U,U] = &\underbrace{\sum_{i=1}^m \left[\lambda(l-1)\langle A_i, \hat{X}\hat{X}^\top - M^*\rangle^{l-2} + 1\right]\langle A_i, U\hat{X}^\top + \hat{X}U^\top\rangle^2}_{B} \\ &+ \underbrace{\sum_{i=1}^m \left[\lambda\langle A_i, \hat{X}\hat{X}^\top - M^*\rangle^{l-1} + \langle A_i, \hat{X}\hat{X}^\top - M^*\rangle\right]\langle A_i, 2UU^\top\rangle}_{A} \quad \forall U \in \mathbb{R}^{n\times r_{search}}\end{aligned} \tag{5.23}$$

Since

$$\langle A_i, U\hat{X}^\top + \hat{X}U^\top\rangle^2 \leq \sum_{i=1}^{m}\langle A_i, U\hat{X}^\top + \hat{X}U^\top\rangle^2 \leq (1+\delta)\|U\hat{X}^\top + \hat{X}U^\top\|_F^2 \quad \forall i$$

then if we choose U *such that*

$$U = uq^\top, \quad \|u\|_2, \|q\|_2 = 1, \quad \|\hat{X}q\|_2 = \sigma_r(\hat{X}),\ u^\top \nabla f^l(\hat{X}\hat{X}^\top)u = \lambda_{\min}(\nabla f_\lambda^l(\hat{X}\hat{X}^\top)),$$

it can be shown that

$$\begin{aligned}(1+\delta)\|U\hat{X}^\top + \hat{X}U^\top\|_F^2 &= (1+\delta)\|u(\hat{X}q)^\top + (\hat{X}q)u^\top\|_F^2 \\ &= 2(1+\delta)\|\hat{X}q\|_F^2 + 2(1+\delta)\left(q^\top(\hat{X}^\top u)\right)^2 \\ &= 2(1+\delta)\sigma_r(\hat{X})^2\end{aligned}$$

since $(\hat{X}q)u^\top = 0$ *as* u *is an eigenvector of* $\nabla f_\lambda^l(\hat{X}\hat{X}^\top)$*, which is orthogonal to* $\hat{X}$ *as required by* (5.10)*. This further implies that*

$$\begin{aligned}B &\leq \sum_{i=1}^{m}\left[\lambda(l-1)\langle A_i, \hat{X}\hat{X}^\top - M^*\rangle^{l-2}\langle A_i, U\hat{X}^\top + \hat{X}U^\top\rangle^2\right] + 2(1+\delta)\sigma_r(\hat{X})^2 \\ &\leq 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)\sum_{i=1}^{m}\langle A_i, \hat{X}\hat{X}^\top - M^*\rangle^{l-2} + 1\right] \\ &= 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)\|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_{l-2}^{l-2} + 1\right] \\ &\leq 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)(\|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_2)^{l-2} + 1\right] \\ &\leq 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)(\sqrt{1+\delta}\|\hat{X}\hat{X}^\top - M^*\|_F)^{l-2} + 1\right] \\ &= 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)(1+\delta)^{(l-2)/2}\|\hat{X}\hat{X}^\top - M^*\|_F^{l-2} + 1\right]\end{aligned} \tag{5.24}$$

Also, given (5.12a) *and our choice of* U*, it is apparent that*

$$A = 2u^\top \nabla f_\lambda^l(\hat{X}\hat{X}^\top)u = 2\lambda_{\min}(\nabla f_\lambda^l(\hat{X}\hat{X}^\top)) \tag{5.25}$$

Therefore, given (5.24)*,* (5.25) *and* (5.22)*, by substituting them back into* (5.23)*, we arrive at*

$$\nabla^2 h^l(\hat{X})[U,U] \leq 2\sigma_r(\hat{X})^2\left(\lambda(l-1)(1+\delta)^{l/2}\|\hat{X}\hat{X}^\top - M^*\|_F^{l-2} + (1+\delta)\right) - 2L/\operatorname{tr}(M^*) \tag{5.26}$$

so the right-hand side of the above equation is strictly negative if (5.6) *is satisfied.*

Part II

The Noisy Regime

Chapter 6

Noisy Matrix Sensing

6.1 Background and Related Work

In this chapter, we turn our attention to matrix sensing problems that are corrupted by noise. As preluded in the introduction section, we aim to solve a problem in this form

$$\min_{X \in \mathbb{R}^{n \times r_{\text{search}}}} h_w(X) := \frac{1}{2}\|\mathcal{A}(XX^T) - \tilde{b}\|^2, \quad \tilde{b} = b - w, \quad w \sim \mathcal{D} \tag{6.1}$$

for the exact definitions of $\mathcal{A}$ and our motivations to solve this problem please refer to Section 1.2 for a more in-depth, detailed review. While the notation $b \in \mathbb{R}^m$ is oftentimes used to represent general measurements, in this work we use b to specifically represent the noiseless, perfect measurements, i.e., $b = \mathcal{A}(M^*)$, where M^* is the ground truth matrix we hope to recover. We model the noise w separately in the objective function of (6.1). Random noise w is assumed to be a general, unknown noise, and our only requirement is that it come from a finite-variance family. From the user's perspective, w is hidden in the measurements and the user can only observe $b - w$, which is the vector of corrupted measurements.

In this chapter, due to the presence of noise, it is impossible to assert that the global solution will remain M^* due to the perturbations. Therefore, we instead quantify the distance between any arbitrary local minimizer $\hat{X}$ and the ground truth in terms of the Frobenius distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ through the lens of restricted isometry property (RIP). In particular, we upper-bound the Frobenius distance using the problem's RIP constant. This work offers the tightest known bound in terms of RIP constant in the noisy and over-parametrized regime. Additionally, although we focus on the symmetric matrix sensing problem, our results can also be applied to the asymmetric problem in which M^* is allowed to be rectangular. This is due to the fact that every asymmetric problem can be equivalently transformed into a symmetric one [24]. For detailed introduction please refer to Section 2.2.

6.1.1 Related works

It is widely known that the RIP condition may guarantee some desirable properties for the geometric optimization landscape of (6.1), namely that is has no spurious local minimizers [24, 5, 65, 103, 106, 101, 93, 6, 27, 107, 95]. Please find introduction to RIP conditions and related concepts in Section 1.2. However, the majority of the existing results have focused on the noiseless, exact-parametrized regime, in which $w = 0$ and $r_{\text{search}} = r$. Although these problems results offer strong theoretical guarantees, they cannot be applied to many real-world problems due to using a simplistic noiseless model and making assumptions on the availability of r. Thus, this work aims to address this issue.

There are some recent papers that specifically study the overparametrized problem (6.1) in the noiseless setting, namely when $w = 0$. In particular, both [48, 76] analyze the convergence behavior of (6.1) with a small initialization. In particular, these papers state that as long as the initialization and step size are small enough, the vanilla gradient descent algorithm applied to noiseless (6.1) will converge to the ground truth solution M^* after a sufficient number of steps. Aside from the lack of consideration of noise, these results differ from our work in two main aspects. First, these results require that the initialization be small in magnitude, while our work studies the global optimization landscape of the problem using the strict saddle property that is agnostic of initialization. Since an initial point with a small magnitude may be far from the true solution, this could seriously affect the convergence time. Therefore, it is important to allow initializing the algorithm at any provided estimate of the solution rather than zero. Second, the RIP bounds needed in both of the above paper scale with $1/\sqrt{r}$, the condition number of M^*, and an unknown constant that could be potentially tiny. These bounds are greatly reduced in our work to only scale with $1/\sqrt{r/r_{\text{search}}}$, which is much better than $1/\sqrt{r}$. On the other hand, [95] developed a sharp bound on the absence of spurious local minima for problem (6.1) in the noiseless case. The paper states that if the measurement operator $\mathcal{A}$ satisfies the $(\delta, r_{\text{search}} + r)$-RIP property with $\delta < 1/(1 + \sqrt{r/r_{\text{search}}})$, then the gradient descent algorithm can be applied to solve the problem with enough iterations. In the exact parametrized case, this simplifies to $\delta < 1/2$, which is a sharp bound due to the counterexample given in [98] that has spurious local minima under $\delta = 1/2$.

When we consider the effect of the noise, the problem becomes highly complicated since the global minimizer X^* of (6.1) does not satisfy the equation $X^*X^{*T} = M^*$ anymore as the noise deteriorates the landscape of (6.1). Therefore, instead of ensuring that there exists no spurious local minima, we study the distance between any local minimizer $\hat{X}$ and the ground truth under the presence of RIP assumption. One important result in the noisy and exact-parametrized regime is as follows:

Theorem 10 *([5], Theorem 3.1) Suppose that* $w \sim \mathcal{N}(0, \sigma_w^2 I_m)$*,* $r_{search} = r$ *and* $\mathcal{A}(\cdot)$ *has the* $(\delta, 4r)$*-RIP property with* $\delta < 1/10$*. Then, with probability at least* $1 - 10/n^2$*, every local minimizer* $\hat{X}$ *of problem* (6.1) *satisfies the inequality*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq 20\sqrt{\frac{\log(n)}{m}}\sigma_w.$$

Theorem 31 in [24] further improves the above result by replacing the $(\delta, 4r)$-RIP property with the $(\delta, 2r)$-RIP property. [47] studies a similar noisy low-rank matrix recovery problem with the l_1 norm.

In the noisy and overparametrized regime, [108] also offers some local guarantees. In that work, the authors proved that if the gradient descent algorithm is initialized close to M^* and that n is large enough, the projection of $XX^\top$ onto the rank-r submanifold is close to M^* after enough steps. Although both the above paper and our work both focus on the noisy and overparametrized regime, there are two major differences. First, the above paper only focuses on sub-Gaussian sensing matrix, whereas our work can be applied to all sensing matrices that satisfy the RIP condition. Although the sub-Gaussian sensing matrix meets the RIP requirement when the number of measurements is large enough, there are many problems satisfying the RIP condition which are not sampled from sub-Gaussian distributions. Second, the abovementioned work only analyzes local convergence, while we will prove the absence of spurious local minima and global strict-saddle property which can ensure a polynomial-time global convergence with an arbitrary initialization.

In the noisy and overparametrized regime the existing literature lacks a global guarantee similar to Theorem 10 to characterize the optimization landscape regardless of initialization. This issue will be addressed in this paper by showing that a major generalization of Theorem 10 holds for the noisy problem in both the exact parametrized and the overparametrized regimes, provided that the same RIP assumption needed for the noiseless problem holds. Table 7.1 briefly summarizes our result compared with the existing literature.

Earlier works such as [107, 31, 7] established the strict saddle property for the noiseless and exact parametrized problem, which essentially states that any matrix whose gradient is small and whose Hessian is almost positive semidefinite (gradient and Hessian of the objective function evaluated at this matrix) must be sufficiently close to a global minimizer. This property, together with certain local regularity property near the ground truth, implies the global linear convergence for the perturbed gradient descent method. In other words, the algorithm will return a solution $\hat{X}$ satisfying $\|\hat{X}\hat{X}^\top - M^*\|_F \leq e$ after $O(\log(1/e))$ number of iterations [31]. In this paper, we prove a similar strict saddle property for the noisy problem in both the exact parametrized and the overparametrized cases. However, in the noisy problem, even if the local search algorithm finds the global minimum, it cannot recover the ground

Paper	Noise	RIP	Parametrization	Initialization
[5]	Gaussian	$\delta < 1/10$	$r_{\text{search}} = r$	Arbitrary
[95]	Noiseless	$\delta < 1/\sqrt{r/r_{\text{search}}}$	$r_{\text{search}} \geq r$	Arbitrary
[48]	Noiseless	$\delta < \mathcal{O}(1/(\kappa^3\sqrt{r}\log^2 n))$	$r_{\text{search}} \geq r$	$\|X_0\|_F \leq \mathcal{O}(1/n)$
[76]	Noiseless	$\delta < \mathcal{O}(1/(\kappa^4\sqrt{r}))$	$r_{\text{search}} \geq r$	$\|X_0\|_F \leq \mathcal{O}(\frac{\|M^*\|_2^2}{\kappa^7 n})$
[108]	General Noise	sub-Guassian $\mathcal{A}$ $m \geq \mathcal{O}(n\log^3 n)$	$r_{\text{search}} \geq r$	$\|X_0X_0^\top - M^*\|_2$ $\leq \mathcal{O}(\sigma_r(M^*))$
Ours	General Noise	$\delta < 1/\sqrt{r/r_{\text{search}}}$	$r_{\text{search}} \geq r$	Arbitrary

Table 6.1: Comparison between our result and the existing literature. $\kappa := \|M^*\|_2/\sigma_r(M^*)$ is defined to be the condition number of the ground truth.

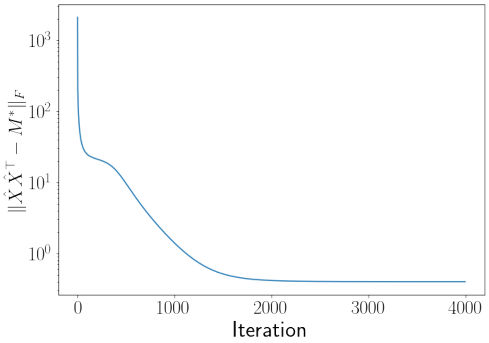


Figure 6.1: The evolution of the error between the found solution $\hat{X}\hat{X}^\top$ and the ground truth M^* during the iterations of the gradient descent method for a noisy problem with the RIP constant $\delta < 1/2$.

truth exactly. As such, it is no longer meaningful to discuss the convergence rate because the error between the found solution and the ground truth has two sources: the difference between $X^*X^{*\top}$ and the ground truth M^* where X^* denotes an exact global minimizer of the problem, and the difference between $\hat{X}\hat{X}^\top$ and $X^*X^{*\top}$ where $\hat{X}$ denotes the approximate solution found by the algorithm. Using our strict saddle property, we can characterize the time point when the errors induced by the above two sources are roughly equal. As demonstrated via an example in Fig. 6.1, it is almost futile to run the algorithm beyond a certain number of iterations since the error will be dominated by the former one after some time.

6.2 Global Optimization Landscape under Noise

We first present the global guarantee on the local minimizers of the problem (6.1). To simplify the notation, we use a matrix representation of the measurement operator $\mathcal{A}$ as follows:

$$\mathbf{A} = [\operatorname{vec}(A_1), \operatorname{vec}(A_2), \dots, \operatorname{vec}(A_m)]^\top \in \mathbb{R}^{m \times n^2}.$$

Then, $\mathbf{A}\operatorname{vec}(M) = \mathcal{A}(M)$ for every matrix $M \in \mathbb{R}^{n \times n}$.

Theorem 11 *Given arbitrary positive integers r_{search} and r such that $r_{search} \geq r$, assume that the linear operator $\mathcal{A}$ satisfies the $(\delta, r_{search} + r)$-*RIP *property with $\delta < 1/(1+\sqrt{r/r_{search}})$. For every $\epsilon > 0$ and every arbitrary local minimizer $\hat{X} \in \mathbb{R}^{n \times r_{search}}$ of problem* (6.1) *satisfying:*

$$\|\nabla h_w(\hat{X})\| = 0, \quad \nabla^2 h_w(\hat{X}) \succeq 0,$$

if $\sigma_{r_{search}}(\hat{X}) \leq \sqrt{\frac{\epsilon}{1+\delta}}$, then

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{\epsilon\sqrt{r_{search}} + \sqrt{\epsilon^2 r_{search} - 16(1-\delta)\epsilon\sqrt{r_{search}}\|M^*\|_F}}{2(1-\delta)}, \tag{6.2}$$

and conversely if $\sigma_{r_{search}}(\hat{X}) > \sqrt{\frac{\epsilon}{1+\delta}}$, then

$$\begin{aligned}\left(\frac{1-\delta}{1+\delta} - \frac{\sqrt{r/r_{search}}}{2+\sqrt{r/r_{search}}}\right)\|\hat{X}\hat{X}^\top - M^*\|_F \leq \\ 2\epsilon\sqrt{r} + 2\sqrt{2\epsilon(1+\delta)}(\|\hat{X}\hat{X}^\top - M^*\|_F^{1/2} + \|M^*\|_F^{1/2}),\end{aligned} \tag{6.3}$$

all with probability at least $\mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon)$ in both cases.

Note that (6.3) is a quadratic inequality of $\|\hat{X}\hat{X}^\top - M^*\|_F^{1/2}$, so it can be easily arranged such that $\|\hat{X}\hat{X}^\top - M^*\|_F^{1/2}$ is on the LHS of the inequality with a constant on the RHS, just as (6.2). The reason that (6.3) is presented in the current form is because the closed-form expression will be overly complicated, and we recommend solving the quadratic inequality with parameters plugged in.

The reason for the existence of two inequalities in Theorem 11 is the split of its proof into two cases. The first case is associated with the r_{search}-th smallest singular value of $\hat{X}$ being small and the second case is the opposite, which are respectively handled by Lemma 24 and Lemma 25.

Theorem 11 is a major extension of the state-of-the-art result stating that the noiseless problem has no spurious local minima under the same assumption of the $(\delta, r_{\text{search}} + r)$-RIP property with $\delta < 1/(1 + \sqrt{r/r_{\text{search}}})$. The reason is that in the case when the noise w is equal to zero, one can choose an arbitrarily small ϵ in Theorem 11 to conclude from the inequalities (6.2) and (6.3) that $\hat{X}\hat{X}^\top = M^*$ for every local minimizer $\hat{X}$. Moreover, when the RIP constant δ further decreases from $1/(1+\sqrt{r/r_{\text{search}}})$, the upper bound on $\|\hat{X}\hat{X}^\top - M^*\|_F$ will also decrease, which means that a local minimizer found by local search methods will be closer to the ground truth M^*. This suggests that the RIP condition is able to not only guarantee the absence of spurious local minima as shown in the previous literature but also mitigate the influence of the noise in the measurements.

Compared with the existing results such as Theorem 10, our new result has two advantages even when specialized to the exact parametrized case $r_{\text{search}} = r$. First, by improving the RIP constant from 1/10 to 1/2, one can apply the results on the location of spurious local minima to a much broader class of problems, which can often help reduce the number of measurements. For example, in the case when the measurements are given by random Gaussian matrices, it is proven in [13] that to achieve the $(\delta, 2r)$-RIP property the minimum number of measurements needed is on the order of $O(1/\delta^2)$. By improving the RIP constant in the bound, we can significantly reduce the number of measurements while still keeping the benign landscape. In applications such as learning for energy networks, there is a fundamental limit on the number of measurements that can be collected due to the physics of the problem [35]. Finding a better bound on RIP helps with addressing the issues with the number of measurements needed to reliably solve the problem. Second, Theorem 10 is just about the probability of having all spurious solutions in a fixed ball around the ground truth of radius $O(\sigma_w)$ instead of balls of arbitrary radii, and this fixed ball could be a large one depending on whether the noise level σ_w is fixed or scales with the problem. On the other hand, in Theorem 11, we consider the probability $\mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon)$ for any arbitrary value of ϵ. By having a flexible ϵ, our work not only improves the RIP constant but also allows computing the probability of having all spurious solutions in any given ball.

In the special case of rank $r_{\text{search}} = r = 1$, the conditions (6.2) and (6.3) in Theorem 11 can be substituted with a simpler condition as presented below.

Theorem 12 *Consider the case $r_{search} = r = 1$ and assume that the linear operator $\mathcal{A}$ satisfies the $(\delta, 2)$-RIP property with $\delta < 1/2$. For every $\epsilon > 0$, with probability at least $\mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon)$, every arbitrary local minimizer $\hat{X} \in \mathbb{R}^{n\times r_{search}}$ of problem (6.1) satisfies*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{3(1+\sqrt{2})\epsilon(1+\delta)}{1-2\delta}. \tag{6.4}$$

In the case when the RIP constant δ is not less than $1/(1+\sqrt{r/r_{\text{search}}})$, it is not possible to achieve a global guarantee similar to Theorem 11 or Theorem 12 since it is known that the problem may have a spurious solution even in the noiseless case [101, 95].Instead, we turn to local guarantees by showing that every arbitrary local minimizer $\hat{X}$ of problem (6.1) is either close to the ground truth M^* or far away from it in terms of the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$.

Theorem 13 *Assume that the linear operator $\mathcal{A}$ satisfies the $(\delta, r_{search} + r)$-RIP property for some $\delta \in [0, 1)$. Consider arbitrary constants $\epsilon > 0$ and $\tau \in (0, 1)$ such that $\delta < \sqrt{1-\tau}$. Every local minimizer $\hat{X} \in \mathbb{R}^{n \times r_{search}}$ of problem (6.1) satisfying*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \tau\lambda_r(M^*) \tag{6.5}$$

will satisfy

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{\epsilon\sqrt{r_{search}} + \sqrt{\epsilon^2 r_{search} - 16(1-\delta)\epsilon\sqrt{r_{search}}\|M^*\|_F}}{2(1-\delta)} \tag{6.6}$$

if $\sigma_{r_{search}}(\hat{X}) \leq \sqrt{\frac{\epsilon}{1+\delta}}$, and will satisfy

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{\sqrt{\epsilon}(1+\delta)^{3/2}C(\tau, M^*)}{\sqrt{1-\tau} - \delta} \tag{6.7}$$

if $\sigma_{r_{search}}(\hat{X}) > \sqrt{\frac{\epsilon}{1+\delta}}$, all with probability at least $\mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon)$ in both cases. Here,

$$C(\tau, M^*) = \sqrt{2(\lambda_1(M^*) + \tau\lambda_r(M^*))}.$$

The upper bounds in (6.5), (6.6) and (6.7) define an outer ball and an inner ball centered at the ground truth M^*. In particular, the radius of the outer ball is the RHS of (6.5), and the radius of the inner ball is either the RHS of (6.6) or (6.7), and for convenience's sake it can be assumed to be the larger of the two. Theorem 13 states that there is no local minimizer in the ring between the two balls, which means that bad local minimizers are located outside the outer ball. Note that the problem could be highly non-convex when δ is close to 1, while this theorem shows a benign landscape in a local neighborhood of the solution. Furthermore, similar to Theorem 11 and Theorem 12, as ϵ approaches zero, the inner ball shrinks to the ground truth. Hence, when the problem is noiseless, Theorem 13 shows that every local minimizer $\hat{X}$ satisfying (6.5) must have $\hat{X}\hat{X}^\top = M^*$. It is true that most well-posed problems will exhibit a benign landscape close to their global minimizer. However, depending on the nature of the problem, this local neighborhood can be arbitrarily small in general. What Theorem 13 does is that we characterize the size of this neighborhood using the RIP constant and noise intensity. In the noiseless and exact parametrized case, this exactly recovers Theorem 5 in [93]. Our theorem significantly generalizes the previous result by showing that the same conclusion also holds in the overparametrized regime.

As a remark, all the theorems in this section are applicable to arbitrary noise models since they make no explicit use of the probability distribution of the noise. The only required information is the probability $\mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon)$, which can be computed or bounded when the probability distribution of the noise is given, as illustrated in Section 6.4. The proof to all Theorems presented in this section can also be found in Appendix 6.A.

6.3 Strict Saddle and Convergence under Noise

The results presented above are all about the locations of the local minimizers. They do not automatically imply the global convergence of local search methods with a fast convergence rate. To provide performance guarantees for local search methods, the next theorem establishes a stronger property for the landscape of the noisy problem that is similar to the strict saddle property in the literature [24], which essentially states that all approximate second-order critical points are close to the ground truth.

Theorem 14 *Given arbitrary positive integers r_{search} and r such that $r_{search} \geq r$, assume that the linear operator $\mathcal{A}$ satisfies the $(\delta, r_{search} + r)$-*RIP *property with $\delta < 1/(1 + \sqrt{r/r_{search}})$. For every $\epsilon > 0$, $\kappa \geq 0$ and every matrix $\hat{X} \in \mathbb{R}^{n \times r_{search}}$ satisfying*

$$\|\nabla h_w(\hat{X})\| \leq \kappa \|\hat{X}\|_2, \quad \nabla^2 h_w(\hat{X}) \succeq -\kappa I_{nr_{search}}, \quad \zeta := \epsilon + \kappa/2 \tag{6.8}$$

if $\sigma_{r_{search}}(\hat{X}) \leq \sqrt{\frac{\zeta}{1+\delta}}$, then

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{\zeta\sqrt{r_{search}} + \sqrt{\zeta^2 r_{search} - 16(1-\delta)\zeta\sqrt{r_{search}}\|M^*\|_F}}{2(1-\delta)}, \tag{6.9}$$

and conversely if $\sigma_{r_{search}}(\hat{X}) > \sqrt{\frac{\zeta}{1+\delta}}$, then

$$\begin{split}\left(\frac{1-\delta}{1+\delta} - \frac{\sqrt{r/r_{search}}}{2 + \sqrt{r/r_{search}}}\right)\|\hat{X}\hat{X}^\top - M^*\|_F \leq \zeta\sqrt{r} \\ + 2\sqrt{\zeta(1+\delta)}(\|\hat{X}\hat{X}^\top - M^*\|_F^{1/2} + \|M^*\|_F^{1/2}),\end{split} \tag{6.10}$$

all with probability at least $\mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon)$ in both cases.

Note that (6.10) is a quadratic inequality of $\|\hat{X}\hat{X}^\top - M^*\|_F^{1/2}$, so it can be easily arranged such that $\|\hat{X}\hat{X}^\top - M^*\|_F^{1/2}$ is on the LHS of the inequality with a constant on the RHS, just as (6.9). The reason that (6.10) is presented in the current form is because the closed-form expression will be overly complicated, and we recommend solving the quadratic inequality with parameters plugged in. By Theorem 14, the error $\|\hat{X}\hat{X}^\top - M^*\|_F$ in both (6.9) and (6.10) is induced by the measurement noise characterized by ϵ, together with the inaccuracy of the local search algorithm captured by κ. $\hat{X}\hat{X}^\top$ will be close to the ground truth if ϵ and κ are relatively small, and the contribution from κ to the bounds is exactly half of that from ϵ. Since ϵ is a constant which cannot be decreased during the iterations, it is reasonable to design an algorithm to find an approximate solution $\hat{X}$ satisfying (6.8) with $\epsilon = \kappa/2$ to strike a balance between the probabilistic lower bound and the required number of iterations. To simplify the analysis, our strict saddle property in Theorem 14 is different from the traditional ones [7, 24] which are usually stated as that $\|\hat{X}\hat{X}^\top - M^*\|_F$ is small if $\hat{X}$ satisfies

$$\|\nabla h_w(\hat{X})\| \leq \tilde{\kappa}, \quad \nabla^2 h_w(\hat{X}) \succeq -\tilde{\kappa} I_{nr_{\text{search}}}, \tag{6.11}$$

for a sufficiently small $\tilde{\kappa} > 0$. In [31], it is proven that the perturbed gradient descent method with an arbitrary initialization will find a solution $\hat{X}$ satisfying (6.11) with a high probability in $O(\text{poly}(1/\tilde{\kappa}))$ iterations. Using the assumption that $r > 0$ and thus $0_{n\times n}$ is not the ground truth, in the proof of the next theorem, we will see that the conditions in (6.11) will imply the ones in (6.8) if $\tilde{\kappa}$ is chosen appropriately. This establishes the global convergence for the noisy low-rank matrix recovery problems in both the exact parametrized regime and the overparametrized regime.

Theorem 15 *Let $D \in (0, 1]$ and $\lambda_0 > 0$ be constants such that*

$$\lambda_{\min}(\nabla^2 h_0(\hat{X})) < -\lambda_0$$

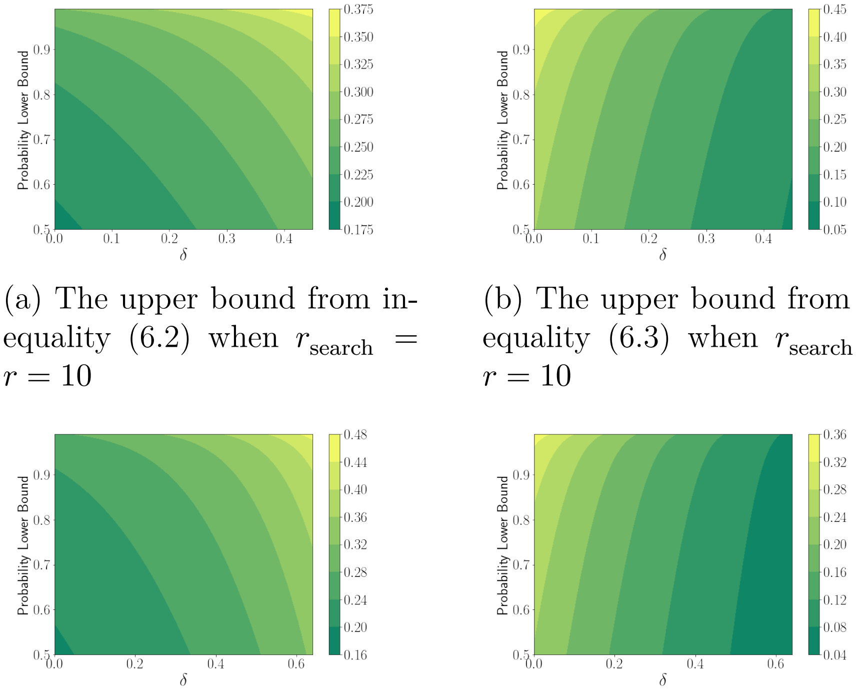
holds for every matrix $\hat{X} \in \mathbb{R}^{n\times r_{search}}$ with $\|\hat{X}\|_2 < D$, where h_0 is the noiseless objective function, i.e. the function h in (6.1) *satisfying $w = 0$. Assume that the linear operator $\mathcal{A}$ satisfies the $(\delta, r_{search} + r)$-*RIP *property with $\delta < 1/(1 + \sqrt{r/r_{search}})$. For every $\epsilon \in (0, \lambda_0)$, the perturbed gradient descent method will find a solution $\hat{X}$ satisfying either of the two inequalities* (6.2) *and* (6.3) *with probability at least $\mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon/2)$ in $O(\text{poly}(1/\epsilon))$ number of iterations.*

Note that the constants D and λ_0 can be directly calculated after the measurement operator $\mathcal{A}$ is explicitly given. As an additional remark, the $O(\text{poly}(1/\epsilon))$ number of iterations does not directly imply that a larger ϵ will lead to fewer iterations, since this mostly describes the rate of scaling with ϵ when ϵ is very small. As per the proof, it can be that $\tilde{\kappa} = \lambda_0 - \epsilon$, meaning that a larger ϵ will lead to more iterations in certain cases. The proof to all Theorems presented in this section can also be found in Appendix 6.A.

6.4 Numerical Illustrations

In the next section, we will empirically study the developed probabilistic guarantees and demonstrate the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ between any local minimizer $\hat{X}$ and the ground truth M^* as well as the value of the RIP constant δ required to be satisfied by the linear operator $\mathcal{A}$, in both the exact parametrized regime and the overparametrized regime.

Before delving into the numerical illustration, note that the probability $\mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon)$ used in our theorems can be exactly calculated as long as the distribution of the



(a) The upper bound from inequality (6.2) when $r_{\text{search}} = r = 10$

(b) The upper bound from inequality (6.3) when $r_{\text{search}} = r = 10$

(c) The upper bound from inequality (6.2) when $r_{\text{search}} = 10$ and $r = 2$

(d) The upper bound from inequality (6.3) when $r_{\text{search}} = 10$ and $r = 2$

Figure 6.2: Comparison of the upper bounds given by Theorem 11 for the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ with $\hat{X}$ being an arbitrary local minimizer

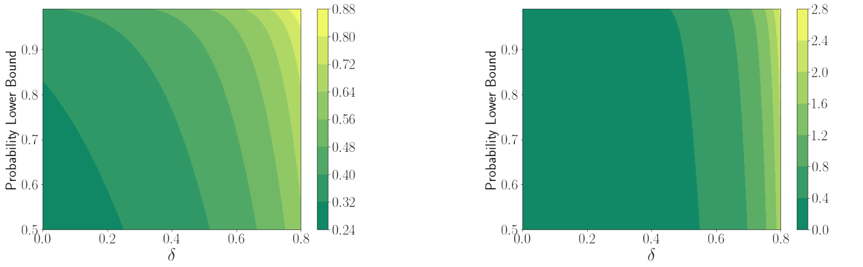
noise w is explicitly given. On the other hand, if we only have partial information for the distribution of w, a lower bound for the probability $\mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon)$ can still be obtained using certain tail bounds. For example, if w is a σ-sub-Gaussian vector, then applying Lemma 1 in [32] leads to

$$1 - 2\mathrm{e}^{-\frac{w_0^2}{16m\sigma^2}} \leq \mathbb{P}(\|w\| \leq w_0) \leq \mathbb{P}(\|\mathbf{A}^\top w\| \leq \epsilon),$$

where $w_0 = \epsilon / \|\mathbf{A}\|_2$.

For numerical illustration, assume that $n = 50$, $m = 10$ and $\|\mathbf{A}\|_2 \leq 2$, while the noise w is a 0.05-sub-Gaussian vector. We also assume that the search rank r_{search} is 10, $\|M^*\|_F = 3.3$, the largest eigenvalue of M^* is 1.5 and its smallest nonzero eigenvalue is 1.

First, we explore the two inequalities (6.2) and (6.3) in Theorem 11 to obtain two upper bounds on $\|\hat{X}\hat{X}^\top - M^*\|_F$, where $\hat{X}$ denotes any arbitrary (worst) local minimizer. For both the exact parametrized case with $r_{\text{search}} = r = 10$ and the overparametrized case with $r_{\text{search}} = 10$ and $r = 2$, Fig. 6.2 gives the contour plots of the two upper bounds on $\|\hat{X}\hat{X}^\top - M^*\|_F$, which hold with the probability given on the y-axis and the RIP constant δ from 0 to $1/(1+\sqrt{r/r_{\text{search}}})$ given on the x-axis. Regardless of the parameterization type, when δ is close to the maximum allowable value $1/(1+\sqrt{r/r_{\text{search}}})$, (6.2) usually dominates the bound, and as δ decreases to 0, (6.3) dominates instead. Furthermore, in the overparametrized regime, (6.3) leads to a tighter bound, while (6.2) remains the same. A similar visualization of the upper bounds given by Theorem 13 for the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ is also presented in Fig. 6.3. We only show the exact parametrized case here because the result is the same for the overparametrized one. It can be observed that (6.7) dominates the bound when δ is closer to 1 and (6.6) dominates when δ is closer to 0.



(a) The upper bound from inequality (6.6) when $r_{\text{search}} = r = 10$

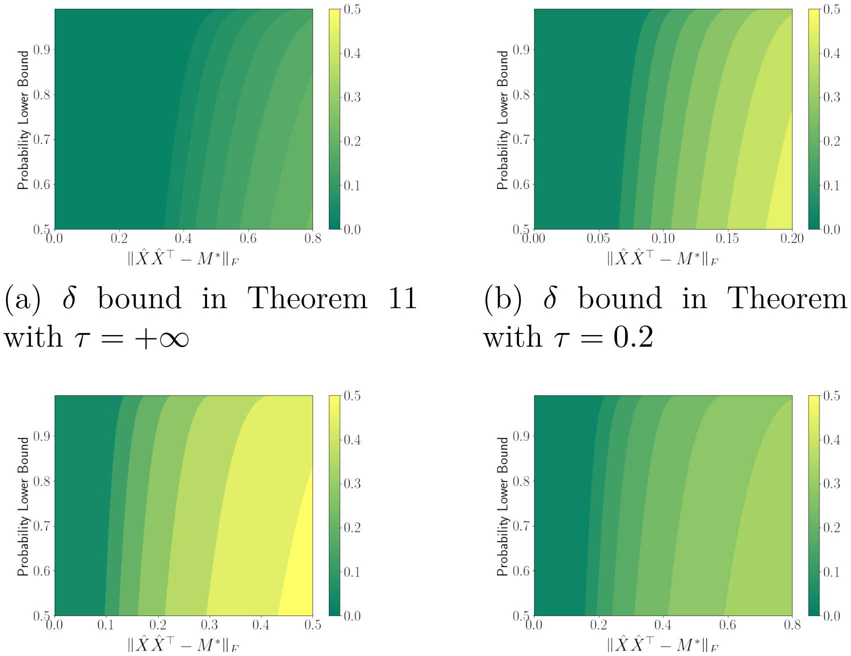
(b) The upper bound from inequality (6.7) when $r_{\text{search}} = r = 10$

Figure 6.3: Comparison of the upper bounds given by Theorem 13 under $\tau = 0.2$ for the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ with $\hat{X}$ being an arbitrary local minimizer

Next, we compare the bounds given by Theorem 11 and Theorem 13. Fig. 7.1 shows the contour plots of the maximum RIP constant δ that is necessary to guarantee that each local minimizer $\hat{X}$ (satisfying the inequality (6.5) when Theorem 13 is applied) lies within a certain neighborhood of the ground truth (measured via the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ on the x-axis) with a given probability on the y-axis, as implied by the respective global and local guarantees. Fig. 7.1 clearly shows how a smaller RIP constant δ leads to a tighter bound on the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ with a higher probability. In addition, the local guarantee generally requires a looser RIP assumption as it still holds even when $\delta > 1/2$. However, as the parameter τ in Theorem 13 increases, the local bound also degrades quickly.

6.5 Summary

In this chapter, we develop global and local analyses for the locations of the local minima of the low-rank matrix recovery problem with noisy linear measurements in both the exact parametrized and the over-parametrized regimes. For the class of



(a) δ bound in Theorem 11 with $\tau = +\infty$

(b) δ bound in Theorem 13 with $\tau = 0.2$

(c) δ bound in Theorem 13 with $\tau = 0.5$

(d) δ bound in Theorem 13 with $\tau = 0.8$

Figure 6.4: Comparison of the maximum RIP constants δ allowed by Theorem 11 and Theorem 13 to guarantee a given maximum distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ for an arbitrary local minimizer $\hat{X}$ satisfying (6.5) with a given probability

noisy problems, regardless of their RIP constants, it is now possible to characterize the worst-case quality of the local minimizers. The major innovation of this work lies in the new proof techniques developed to deal with the over-parameterization and the handling of the random noise via an easy-to-compute concentration bound. Unlike the existing results, the guarantees in our results are distribution-agnostic, meaning that the distribution can be unknown as long as the concentration bound is possible to obtain. The developed results encompass the state-of-the-art results on the non-existence of spurious solutions in the noiseless case. Last but not least, we prove a certain form of the strict saddle property, which guarantees the global convergence of the perturbed gradient descent method in polynomial time regardless of parameterization. Our analyses show how the value of the RIP constant and the intensity of noise affect the landscape of the non-convex learning problem and the locations of the local minima relative to the ground truth.

6.A Missing Details of Chapter 6

Before presenting the proofs, we first compute the gradient and the Hessian of the objective function $f(\hat{X})$ of the problem (6.1):

$$\nabla h_w(\hat{X}) = \hat{\mathbf{X}}^\top \mathbf{A}^\top(\mathbf{A}\mathbf{e} + w),$$
$$\nabla^2 h_w(\hat{X}) = 2I_{r_{\text{search}}} \otimes \operatorname{mat}_S(\mathbf{A}^\top(\mathbf{A}\mathbf{e} + w)) + \hat{\mathbf{X}}^\top \mathbf{A}^\top \mathbf{A}\hat{\mathbf{X}},$$

where

$$\mathbf{e} = \operatorname{vec}(\hat{X}\hat{X}^\top - M^*),$$

and $\hat{\mathbf{X}} \in \mathbb{R}^{n^2 \times nr_{\text{search}}}$ is the matrix satisfying

$$\hat{\mathbf{X}} \operatorname{vec}(U) = \operatorname{vec}(\hat{X}U^\top + U\hat{X}^\top), \quad \forall U \in \mathbb{R}^{n \times r_{\text{search}}}.$$

The first step in the proofs of Theorem 11 and Theorem 14 is to derive necessary conditions for a matrix $\hat{X} \in \mathbb{R}^{n \times r_{\text{search}}}$ to be an approximate second-order critical point, which depend on the linear operator $\mathcal{A}$, the noise $w \in \mathbb{R}^m$, the solution $\hat{X}$, and the parameter κ characterizing how close $\hat{X}$ is to a true second-order critical point.

Lemma 23 *Given $\kappa \geq 0$, assume that $\hat{X} \in \mathbb{R}^{n \times r_{search}}$ satisfies*

$$\|\nabla h_w(\hat{X})\| \leq \kappa \|\hat{X}\|_2, \quad \nabla^2 h_w(\hat{X}) \succeq -\kappa I_{nr_{search}}.$$

Then, it must also satisfy the following inequalities:

$$\|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq (2\|\mathbf{A}^\top w\| + \kappa)\|\hat{X}\|_2, \tag{6.12a}$$
$$2I_{r_{search}} \otimes \operatorname{mat}_S(\mathbf{H}\mathbf{e}) + \hat{\mathbf{X}}^\top \mathbf{H}\hat{\mathbf{X}} \succeq -(2\|\mathbf{A}^\top w\| + \kappa)I_{nr_{search}}, \tag{6.12b}$$

where $\mathbf{H} = \mathbf{A}^\top \mathbf{A}$.

Proof 26 (Proof of Lemma 23) *To obtain condition* (6.12a)*, notice that $\|\nabla h_w(\hat{X})\| \leq \kappa \|\hat{X}\|_2$ implies that*

$$\|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq \|\hat{\mathbf{X}}^\top \mathbf{A}^\top w\| + \|\nabla h_w(\hat{X})\| \leq \|\hat{\mathbf{X}}\|_2 \|\mathbf{A}^\top w\| + \kappa\|\hat{X}\|_2 \leq (2\|\mathbf{A}^\top w\| + \kappa)\|\hat{X}\|_2,$$

in which the last inequality is due to

$$\|\hat{\mathbf{X}} \operatorname{vec}(U)\| = \|\hat{X}U^\top + U\hat{X}^\top\|_F \leq 2\|\hat{X}\|_2 \|U\|_F,$$

for every $U \in \mathbb{R}^{n \times r_{search}}$. Similarly, $\nabla^2 h_w(\hat{X}) \succeq -\kappa I_{nr_{search}}$ implies that

$$2I_{r_{search}} \otimes \operatorname{mat}_S(\mathbf{H}\mathbf{e}) + \hat{\mathbf{X}}^\top \mathbf{H}\hat{\mathbf{X}} \succeq -2I_{r_{search}} \otimes \operatorname{mat}_S(\mathbf{A}^\top w) - \kappa I_{nr_{search}}.$$

On the other hand, the eigenvalues of $I_{r_{search}} \otimes \operatorname{mat}_S(\mathbf{A}^\top w)$ are the same as those of $\operatorname{mat}_S(\mathbf{A}^\top w)$, and each eigenvalue $\lambda_i(\operatorname{mat}_S(\mathbf{A}^\top w))$ of the latter matrix further satisfies

$$|\lambda_i(\operatorname{mat}_S(\mathbf{A}^\top w))| \leq \|\operatorname{mat}_S(\mathbf{A}^\top w)\|_F \leq \|\mathbf{A}^\top w\|,$$

which proves condition (6.12b)*.*

If $\hat{X}$ is a local minimizer of the problem (6.1), Lemma 23 shows that $\hat{X}$ satisfies the inequalities (6.12a) and (6.12b) with $\kappa = 0$. Similarly, Theorem 11 can also be regarded as a special case of Theorem 14 with $\kappa = 0$. The proofs of these two theorems consist of inspecting two cases. The following lemma deals with the first case in which $\hat{X}$ is an approximate second-order critical point with $\sigma_{r_{\text{search}}}(\hat{X})$ being close to zero.

Lemma 24 *Assume that the linear operator $\mathcal{A}$ satisfies the $(\delta, r_{search} + r)$-RIP property. Given $\hat{X} \in \mathbb{R}^{n \times r_{search}}$ and arbitrary constants $\epsilon > 0$ and $\kappa \geq 0$, the inequalities*

$$\sigma_{r_{search}}(\hat{X}) \leq \sqrt{\frac{\epsilon + \kappa/2}{1+\delta}}, \quad \|\nabla h_w(\hat{X})\| \leq \kappa \|\hat{X}\|_2, \quad \nabla^2 h_w(\hat{X}) \succeq -\kappa I_{nr_{search}}$$

and $\|\mathbf{A}^\top w\| \leq \epsilon$ will together imply the inequality (6.9).

Proof 27 (Proof of Lemma 24) *Let $G = \operatorname{mat}_S(\mathbf{He})$ and $u \in \mathbb{R}^n$ be a unit eigenvector of G corresponding to its minimum eigenvalue, i.e.,*

$$\|u\| = 1, \quad Gu = \lambda_{\min}(G)u.$$

In addition, let $v \in \mathbb{R}^r$ be a singular vector of $\hat{X}$ such that

$$\|v\| = 1, \quad \|\hat{X}v\| = \sigma_{r_{search}}(\hat{X}).$$

Let $\mathbf{U} = \operatorname{vec}(uv^\top)$. Then, $\|\mathbf{U}\| \leq 1$ and (6.12b) imply that

$$\begin{aligned} -2\epsilon - \kappa &\leq 2\mathbf{U}^\top (I_{r_{search}} \otimes \operatorname{mat}_S(\mathbf{He}))\mathbf{U} + \mathbf{U}^\top \hat{\mathbf{X}}^\top \mathbf{H} \hat{\mathbf{X}} \mathbf{U} \\ &\leq 2\operatorname{tr}(vu^\top G u v^\top) + (1+\delta)\|\hat{X}vu^\top + uv^\top \hat{X}^\top\|_F^2 \\ &\leq 2\lambda_{\min}(G) + 4(1+\delta)\sigma_{r_{search}}(\hat{X})^2 \\ &\leq 2\lambda_{\min}(G) + 4\epsilon + 2\kappa. \end{aligned} \tag{6.13}$$

On the other hand,

$$\begin{aligned} (1-\delta)\|\hat{X}\hat{X}^\top - M^*\|_F^2 &\leq \mathbf{e}^\top \mathbf{He} = \operatorname{vec}(\hat{X}\hat{X}^\top)^\top \mathbf{He} - \operatorname{vec}(M^*)^\top \mathbf{He} \\ &= \frac{1}{2}\operatorname{vec}(\hat{X})^\top \hat{\mathbf{X}}^\top \mathbf{He} - \langle M^*, \operatorname{mat}_S(\mathbf{He})\rangle \\ &\leq \frac{1}{2}\|\hat{X}\|_F \|\hat{\mathbf{X}}^\top \mathbf{He}\| + \left(3\epsilon + \frac{3\kappa}{2}\right)\operatorname{tr}(M^*) \\ &\leq \left(\epsilon + \frac{\kappa}{2}\right)\|\hat{X}\|_F^2 + \left(3\epsilon + \frac{3\kappa}{2}\right)\operatorname{tr}(M^*), \end{aligned}$$

in which the second last inequality is due to (6.13). In particular, we have that $-\langle M^*, G\rangle = \langle -G, M^*\rangle \leq \lambda_{\max}(-G)\|M^*\|_*$, which gives the desired inequality given

$\lambda_{\max}(-G) = -\lambda_{\min}(G)$ and $\|M^*\|_* = \operatorname{tr}(M^*)$. The last inequality is due to (6.12a). *Furthermore, the right-hand side of the above inequality can be relaxed as*

$$
\begin{aligned}
&\left(\epsilon + \frac{\kappa}{2}\right) \|\hat{X}\|_F^2 + \left(3\epsilon + \frac{3\kappa}{2}\right) \operatorname{tr}(M^*) \\
&\leq \left(\epsilon + \frac{\kappa}{2}\right) \sqrt{r_{search}} \|\hat{X}\hat{X}^\top\|_F + \left(3\epsilon + \frac{3\kappa}{2}\right) \sqrt{r} \|M^*\|_F \\
&\leq \left(\epsilon + \frac{\kappa}{2}\right) \sqrt{r_{search}} \|\hat{X}\hat{X}^\top - M^*\|_F + \left(\epsilon + \frac{\kappa}{2}\right) \sqrt{r_{search}} \|M^*\|_F + \left(3\epsilon + \frac{3\kappa}{2}\right) \sqrt{r} \|M^*\|_F \\
&= \left(\epsilon + \frac{\kappa}{2}\right) \sqrt{r_{search}} \|\hat{X}\hat{X}^\top - M^*\|_F + (4\epsilon + 2\kappa) \sqrt{r_{search}} \|M^*\|_F,
\end{aligned}
$$

which leads to the inequality (6.9).

The remaining case with

$$\sigma_{r_{\text{search}}}(\hat{X}) > \sqrt{\frac{\epsilon + \kappa/2}{1 + \delta}}$$

will be handled in the following lemma using a different method.

Lemma 25 *Assume that the linear operator $\mathcal{A}$ satisfies the $(\delta, r_{search} + r)$-*RIP *property with $\delta < 1/(1 + \sqrt{r/r_{search}})$. Given $\hat{X} \in \mathbb{R}^{n \times r_{search}}$ and arbitrary constants $\epsilon > 0$ and $\kappa \geq 0$, the inequalities*

$$\sigma_{r_{search}}(\hat{X}) > \sqrt{\frac{\epsilon + \kappa/2}{1 + \delta}}, \quad \|\nabla h_w(\hat{X})\| \leq \kappa \|\hat{X}\|_2, \quad \nabla^2 h_w(\hat{X}) \succeq -\kappa I_{nr_{search}}$$

and $\|\mathbf{A}^\top w\| \leq \epsilon$ will together imply the inequality (6.10).

The proofs of both Lemma 25 and the local guarantee in Theorem 13 later generalize the proof of the absence of spurious local minima for the noiseless problem in [101]. Our innovation here is to develop new techniques to analyze approximate optimality conditions for the solutions because unlike the noiseless problem the local minimizers of the noisy one are only approximate second-order critical points of the distance function $\|\mathcal{A}(XX^\top) - b\|^2$. For a fixed matrix $\hat{X}$ and fixed constants $\epsilon > 0$ and $\kappa \geq 0$, one can find an operator $\hat{\mathcal{A}}$ satisfying the $(\delta, r_{\text{search}} + r)$-RIP property with the smallest possible δ such that $\hat{X}$ and $\hat{\mathcal{A}}$ satisfy the conditions (6.12a) and (6.12b) stated in Lemma 23 (with $\|\mathbf{A}^\top w\|$ replaced by ϵ). Let $\delta^*(\hat{X})$ be the RIP constant of the found measurement operator $\hat{\mathcal{A}}$ in this worst-case scenario. Then, for any instance of the problem (6.1) with the linear operator $\mathcal{A}$ satisfying the $(\delta, r_{\text{search}} + r)$-RIP property and the noise w satisfying $\|\mathbf{A}^\top w\| \leq \epsilon$, if $\hat{X}$ is an approximate second-order critical point to this instance satisfying the conditions listed in Lemma 25, we can ensure that $\delta \geq \delta^*(\hat{X})$ by the definition of $\delta^*(\hat{X})$. As we will elaborate in the proof below, since $\delta^*(\hat{X})$ can be lower-bounded analytically in terms of $\|\hat{X}\hat{X}^\top - M^*\|_F$, this will lead to an upper bound on the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$.

Proof 28 (Proof of Lemma 25) *The variable $\delta^*(\hat{X})$ defined above is the optimal value of the following optimization problem:*

$$\begin{aligned}\min_{\delta,\hat{\mathbf{H}}} \quad & \delta \\ \textit{s.t.} \quad & \|\hat{\mathbf{X}}^\top \hat{\mathbf{H}}\mathbf{e}\| \le (2\epsilon+\kappa)\|\hat{X}\|_2, \\ & 2I_{r_{search}} \otimes \operatorname{mat}_S(\hat{\mathbf{H}}\mathbf{e}) + \hat{\mathbf{X}}^\top \hat{\mathbf{H}}\hat{\mathbf{X}} \succeq -(2\epsilon+\kappa) I_{nr_{search}}, \\ & \hat{\mathbf{H}} \textit{ is symmetric and satisfies the } (\delta, r_{search}+r)\text{-RIP} \textit{ property.}\end{aligned} \tag{6.14}$$

Here, a matrix $\hat{\mathbf{H}} \in \mathbb{R}^{n^2\times n^2}$ is said to satisfy the $(\delta, r_{search}+r)$-RIP property if

$$(1-\delta)\|\mathbf{U}\|^2 \le \mathbf{U}^\top \hat{\mathbf{H}}\mathbf{U} \le (1+\delta)\|\mathbf{U}\|^2$$

holds for every matrix $U \in \mathbb{R}^{n\times n}$ with $\operatorname{rank}(U) \le r_{search}+r$ and $\mathbf{U} = \operatorname{vec}(U)$. Obviously, for a linear operator $\hat{\mathcal{A}}$, $\hat{\mathbf{H}} = \hat{\mathbf{A}}^\top\hat{\mathbf{A}}$ satisfies the $(\delta, r_{search}+r)$-RIP property defined above if and only if $\hat{\mathbf{A}}$ satisfies the $(\delta, r_{search}+r)$-RIP property. By the discussion above, we have $\delta \ge \delta^*(\hat{X})$.

However, since the problem (6.14) is non-convex due to the RIP constraint, we may not be able to solve for $\delta^*(\hat{X})$ exactly, and therefore we provide a lower bound instead. To achieve this goal, we introduce the following lemma:

Lemma 26 *If $\eta^*(\hat{X})$ is the optimal value of the following convex optimization problem:*

$$\begin{aligned}\max_{\eta,\tilde{\mathbf{H}}} \quad & \eta \\ \text{s. t.} \quad & \|\tilde{\mathbf{H}}^\top \tilde{\mathbf{H}}\mathbf{e}\| \le (2\epsilon+\kappa)\|\hat{X}\|_2, \\ & 2I_{r_{search}} \otimes \operatorname{mat}_S(\tilde{\mathbf{H}}\mathbf{e}) + \tilde{\mathbf{H}}^\top \tilde{\mathbf{H}} \succeq -(2\epsilon+\kappa) I_{nr_{search}}, \\ & \eta I_{n^2} \preceq \tilde{\mathbf{H}} \preceq I_{n^2},\end{aligned} \tag{6.15}$$

then

$$\frac{1-\eta^*(\hat{X})}{1+\eta^*(\hat{X})} \le \delta^*(\hat{X}) \le \delta, \tag{6.16}$$

where $\delta^*(\hat{X})$ is the optimal value of the problem (6.14).

Therefore, the above lemma directly implies that

$$\eta^*(\hat{X}) \ge \frac{1-\delta^*(\hat{X})}{1+\delta^*(\hat{X})} \ge \frac{1-\delta}{1+\delta}. \tag{6.17}$$

Furthermore, since $\eta^*(\hat{X})$ is the optimal value of the problem (6.15), a semidefinite program in a rather simple form, we could further derive an analytical upper bound for $\eta^*(\hat{X})$ using its Lagrangian dual, which is embodied in the following lemma:

Lemma 27 *The optimal value of* (6.15), $\eta^*(\hat{X})$, *satisfies*

$$\eta^*(\hat{X}) \leq \frac{\sqrt{r/r_{search}}}{2+\sqrt{r/r_{search}}} + \frac{(2\epsilon+\kappa)\sqrt{r}+2\sqrt{(2\epsilon+\kappa)(1+\delta)}\|\hat{X}\|_2}{\|\mathbf{e}\|}. \tag{6.18}$$

Finally, our desired inequality (6.10) *is a direct consequence of* (6.17), (6.18) *and the inequality*

$$\|\hat{X}\|_2 \leq \|\hat{X}\hat{X}^\top\|_F^{1/2} \leq \|\hat{X}\hat{X}^\top - M^*\|_F^{1/2} + \|M^*\|_F^{1/2}.$$

With the proofs of Lemma 26 and Lemma 27 given below, this completes the proof of this lemma.

Proof 29 (Proof of Lemma 26) *Consider a convex reformulation of* (6.14) *by enforcing the RIP constraint over all ranks:*

$$\begin{aligned} \min_{\delta,\hat{\mathbf{H}}} \quad & \delta \\ \text{s.t.} \quad & \|\hat{\mathbf{X}}^\top\hat{\mathbf{H}}\mathbf{e}\| \leq (2\epsilon+\kappa)\|\hat{X}\|_2, \\ & 2I_{r_{search}} \otimes \operatorname{mat}_S(\hat{\mathbf{H}}\mathbf{e}) + \hat{\mathbf{X}}^\top\hat{\mathbf{H}}\hat{\mathbf{X}} \succeq -(2\epsilon+\kappa)I_{nr_{search}}, \\ & (1-\delta)I_{n^2} \preceq \hat{\mathbf{H}} \preceq (1+\delta)I_{n^2}. \end{aligned} \tag{6.19}$$

Lemma 14 in [6] proves that the problem (6.14) *and the problem* (6.19) *have the same optimal value. The remaining step is to solve the optimization problem* (6.19) *for given* $\hat{X}$, ϵ *and* κ. *To make the ensuing proof easier, we aim to further rewrite* (6.19) *by simplifying its second and third constraints. First, we convert the last constraint of* (6.19) *to* $(1-\delta)I_{n^2}/(1+\delta) \preceq \hat{\mathbf{H}}/(1+\delta) \preceq I_{n^2}$, *and realize that one could replace* $(1-\delta)/(1+\delta)$ *with another variable* η, *with a smaller* δ *leading to a bigger* η *and vice versa, and replace* $\hat{\mathbf{H}}$ *with* $\tilde{\mathbf{H}} = \hat{\mathbf{H}}/(1+\delta)$.

Therefore, instead of optimizing for the smallest possible δ, *we can now optimize for the largest possible* η. *To exactly satisfy the second and third constraints after the change of variables, we introduce* (6.15) *and denote its optimal value as* $\eta^*(\hat{X})$, *which is repeated here for the sake of convenience:*

$$\begin{aligned} \max_{\eta,\tilde{\mathbf{H}}} \quad & \eta \\ \text{s.t.} \quad & \|\tilde{\mathbf{H}}^\top\tilde{\mathbf{H}}\mathbf{e}\| \leq (2\epsilon+\kappa)\|\hat{X}\|_2, \\ & 2I_{r_{search}} \otimes \operatorname{mat}_S(\tilde{\mathbf{H}}\mathbf{e}) + \hat{\mathbf{X}}^\top\hat{\mathbf{X}} \succeq -(2\epsilon+\kappa)I_{nr_{search}}, \\ & \eta I_{n^2} \preceq \tilde{\mathbf{H}} \preceq I_{n^2}. \end{aligned}$$

Observe that given any feasible solution $(\delta, \hat{\mathbf{H}})$ *to* (6.19), *the tuple*

$$\left(\eta, \tilde{\mathbf{H}}\right) = \left(\frac{1-\delta}{1+\delta}, \frac{1}{1+\delta}\hat{\mathbf{H}}\right)$$

is a feasible solution to the problem (6.15)*. We prove this statement by checking all the three constraints one by one. For the first constraint,*

$$\|\tilde{\mathbf{H}}^\top \tilde{\mathbf{H}}\mathbf{e}\| = \frac{1}{(1+\delta)^2}\|\hat{\mathbf{X}}^\top \hat{\mathbf{H}}\mathbf{e}\| \le \frac{1}{(1+\delta)^2}(2\epsilon+\kappa)\|\hat{X}\|_2 \le (2\epsilon+\kappa)\|\hat{X}\|_2.$$

For the second constraint,

$$\begin{aligned} 2I_{r_{search}} \otimes \mathrm{mat}_S(\tilde{\mathbf{H}}\mathbf{e}) + \hat{\mathbf{X}}^\top\hat{\mathbf{X}} &\succeq \frac{2}{1+\delta} I_{r_{search}} \otimes \mathrm{mat}_S(\hat{\mathbf{H}}\mathbf{e}) + \frac{1}{1+\delta}\hat{\mathbf{X}}^\top \hat{\mathbf{H}}\hat{\mathbf{X}} \\ &\succeq -\frac{2\epsilon+\kappa}{1+\delta} I_{nr_{search}} \succeq -(2\epsilon+\kappa) I_{nr_{search}}. \end{aligned}$$

Finally, the third constraint is satisfied automatically. Now, since $(1-\delta^*(\hat{X}))/(1+\delta^*(\hat{X}))$ *is a feasible value for* (6.15)*, we have*

$$\eta^*(\hat{X}) \ge \frac{1-\delta^*(\hat{X})}{1+\delta^*(\hat{X})},$$

which further implies (6.16)*.*

Proof 30 (Proof of Lemma 27) *The proof of the upper bound* (6.18) *can be completed by finding a feasible solution to the dual problem of* (6.15)*:*

$$\begin{aligned} \min_{\substack{U_1,U_2,W,\\ G,\lambda,y}} \quad & \mathrm{tr}(U_2) + \langle \hat{\mathbf{X}}^\top \hat{\mathbf{X}}, W\rangle + (2\epsilon+\kappa)\,\mathrm{tr}(W) + (2\epsilon+\kappa)^2 \|\hat{X}\|_2^2 \lambda + \mathrm{tr}(G) \\ \text{s.t.} \quad & \mathrm{tr}(U_1) = 1, \\ & (\hat{\mathbf{X}}y - w)\mathbf{e}^\top + \mathbf{e}(\hat{\mathbf{X}}y - w)^\top = U_1 - U_2, \\ & \begin{bmatrix} G & -y \\ -y^\top & \lambda \end{bmatrix} \succeq 0, \\ & U_1 \succeq 0, \quad U_2 \succeq 0, \quad W = \begin{bmatrix} W_{1,1} & \cdots & W_{r,1}^\top \\ \vdots & \ddots & \vdots \\ W_{r,1} & \cdots & W_{r,r} \end{bmatrix} \succeq 0, \\ & w = \sum_{i=1}^r \mathrm{vec}(W_{i,i}). \end{aligned} \tag{6.20}$$

Before describing the choice of the dual feasible solution, we need to represent the error vector $\mathbf{e}$ *in a different form. Let* $\mathcal{P} \in \mathbb{R}^{n\times n}$ *be the orthogonal projection matrix onto the range of* $\hat{X}$*, and* $\mathcal{P}_\perp \in \mathbb{R}^{n\times n}$ *be the orthogonal projection matrix onto the orthogonal complement of the range of* $\hat{X}$*. Furthermore, let* $Z \in \mathbb{R}^{n\times r_{search}}$ *be a matrix satisfying* $ZZ^\top = M^*$*. Then,* $\mathrm{rank}(Z) = r$*, and* Z *can be decomposed as* $Z = \mathcal{P}Z + \mathcal{P}_\perp Z$*, so there exists a matrix* $R \in \mathbb{R}^{r\times r}$ *such that* $\mathcal{P}Z = \hat{X}R$*. Note that*

$$ZZ^\top = \mathcal{P}ZZ^\top\mathcal{P} + \mathcal{P}ZZ^\top\mathcal{P}_\perp + \mathcal{P}_\perp ZZ^\top \mathcal{P} + \mathcal{P}_\perp ZZ^\top \mathcal{P}_\perp.$$

Thus, if we choose

$$\hat{Y} = \frac{1}{2}\hat{X} - \frac{1}{2}\hat{X}RR^\top - \mathcal{P}_\perp ZR^\top, \quad \hat{y} = \operatorname{vec}(\hat{Y}), \tag{6.21}$$

then it can be verified that

$$\begin{aligned} \hat{X}\hat{Y}^\top + \hat{Y}\hat{X}^\top - \mathcal{P}_\perp ZZ^\top \mathcal{P}_\perp = \hat{X}\hat{X}^\top - ZZ^\top, \\ \langle \hat{X}\hat{Y}^\top + \hat{Y}\hat{X}^\top, \mathcal{P}_\perp ZZ^\top \mathcal{P}_\perp \rangle = 0. \end{aligned} \tag{6.22}$$

Moreover, we have

$$\begin{aligned} \|\hat{X}\hat{Y}^\top + \hat{Y}\hat{X}^\top\|_F^2 &= 2\operatorname{tr}(\hat{X}^\top\hat{X}\hat{Y}^\top\hat{Y}) + \operatorname{tr}(\hat{X}^\top\hat{Y}\hat{X}^\top\hat{Y}) + \operatorname{tr}(\hat{Y}^\top\hat{X}\hat{Y}^\top\hat{X}) \\ &\geq 2\operatorname{tr}(\hat{X}^\top\hat{X}\hat{Y}^\top\hat{Y}) \geq 2\sigma_{r_{search}}(\hat{X})^2\|\hat{Y}\|_F^2, \end{aligned} \tag{6.23}$$

in which the first inequality is due to

$$\operatorname{tr}(\hat{X}^\top\hat{Y}\hat{X}^\top\hat{Y}) = \frac{1}{4}\operatorname{tr}((\hat{X}^\top\hat{X}(I_{r_{search}} - RR^\top))^2) = \frac{1}{4}\operatorname{tr}((\hat{X}(I_{r_{search}} - RR^\top)\hat{X}^\top)^2) \geq 0.$$

Assume first that $Z_\perp = \mathcal{P}_\perp Z \neq 0$. The other case will be handled at the end of this proof. In the case when $Z_\perp \neq 0$, we also have $\hat{X}\hat{Y}^\top + \hat{Y}\hat{X}^\top \neq 0$. Otherwise, the inequality (6.23) *and the assumption $\sigma_{r_{search}}(\hat{X}) > 0$ imply that $\hat{Y} = 0$. The orthogonality and the definition of $\hat{Y}$ in* (6.21) *then give rise to*

$$\hat{X} - \hat{X}RR^\top = 0, \quad \mathcal{P}_\perp ZR^\top = 0.$$

The first equation above implies that r_{search} is invertible since $\hat{X}$ has full column rank, which contradicts $Z_\perp \neq 0$. Now, define the unit vectors

$$\hat{u}_1 = \frac{\hat{\mathbf{X}}\hat{y}}{\|\hat{\mathbf{X}}\hat{y}\|}, \quad \hat{u}_2 = \frac{\operatorname{vec}(Z_\perp Z_\perp^\top)}{\|Z_\perp Z_\perp^\top\|_F}.$$

Then, $\hat{u}_1 \perp \hat{u}_2$ and

$$\mathbf{e} = \|\mathbf{e}\|(\sqrt{1-\alpha^2}\hat{u}_1 - \alpha\hat{u}_2) \tag{6.24}$$

with

$$\alpha = \frac{\|Z_\perp Z_\perp^\top\|_F}{\|\hat{X}\hat{X}^\top - ZZ^\top\|_F}. \tag{6.25}$$

We first describe our choices of the dual variables W and y (which will be rescaled later). Let

$$\hat{X}^\top\hat{X} = QSQ^\top, \quad Z_\perp Z_\perp^\top = PGP^\top,$$

with orthogonal matrices Q, P and diagonal matrices S, G such that $S_{11} = \sigma_{r_{search}}(\hat{X})^2$. Fix a constant $\gamma \in [0,1]$ that is to be determined and define

$$V_i = k^{1/2} G_{ii}^{1/2} P E_{i1} Q^\top, \quad \forall i \in \{1, \dots, r\},$$
$$W = \sum_{i=1}^r \operatorname{vec}(V_i) \operatorname{vec}(V_i)^\top, \quad y = l\hat{y},$$

with $\hat{y}$ defined in (6.21) *and*

$$k = \frac{\gamma}{\|\mathbf{e}\| \|Z_\perp Z_\perp^\top\|_F}, \quad l = \frac{\sqrt{1-\gamma^2}}{\|\mathbf{e}\| \|\hat{\mathbf{X}}\hat{y}\|}.$$

Here, E_{ij} is the elementary matrix of size $n \times r_{search}$ with the (i,j)-entry being 1. *By our construction, $\hat{X}^\top V_i = 0$, which implies that*

$$\begin{aligned}\langle \hat{\mathbf{X}}^\top \hat{\mathbf{X}}, W \rangle &= \sum_{i=1}^r \|\hat{X} V_i^\top + V_i \hat{X}^\top\|_F^2 = 2 \sum_{i=1}^r \operatorname{tr}(\hat{X}^\top \hat{X} V_i^\top V_i) \\ &= 2k\sigma_{r_{search}}(\hat{X})^2 \sum_{i=1}^r G_{ii} = 2\beta\gamma,\end{aligned} \tag{6.26}$$

with

$$\beta = \frac{\sigma_{r_{search}}(\hat{X})^2 \operatorname{tr}(Z_\perp Z_\perp^\top)}{\|\hat{X}\hat{X}^\top - ZZ^\top\|_F \|Z_\perp Z_\perp^\top\|_F}. \tag{6.27}$$

In addition,

$$\operatorname{tr}(W) = \sum_{i=1}^r \|V_i\|_F^2 = k \sum_{i=1}^r G_{ii} = k \operatorname{tr}(Z_\perp Z_\perp^\top) \le \frac{\sqrt{r}}{\|\mathbf{e}\|}, \tag{6.28}$$

and

$$w = \sum_{i=1}^r \operatorname{vec}(W_{i,i}) = \sum_{i=1}^r V_i V_i^\top = k Z_\perp Z_\perp^\top.$$

Therefore,

$$\hat{\mathbf{X}} y - w = \frac{1}{\|\mathbf{e}\|} (\sqrt{1-\gamma^2} \hat{u}_1 - \gamma \hat{u}_2),$$

which together with (6.24) *implies that*

$$\|\mathbf{e}\| \|\hat{\mathbf{X}} y - w\| = 1, \quad \langle \mathbf{e}, \hat{\mathbf{X}} y - w \rangle = \gamma\alpha + \sqrt{1-\gamma^2}\sqrt{1-\alpha^2} = \psi(\gamma). \tag{6.29}$$

Next, the inequality (6.23) *and the assumption on $\sigma_{r_{search}}(\hat{X})$ imply that*

$$(4\epsilon + 2\kappa)\|y\| \le \frac{\sqrt{1-\gamma^2}(4\epsilon + 2\kappa)}{\sqrt{2}\sigma_{r_{search}}(\hat{X})\|\mathbf{e}\|} \le \frac{2\sqrt{(2\epsilon+\kappa)(1+\delta)}}{\|\mathbf{e}\|}. \tag{6.30}$$

Define

$$M = (\hat{\mathbf{X}}y - w)\mathbf{e}^\top + \mathbf{e}(\hat{\mathbf{X}}y - w)^\top,$$

and decompose M as $M = [M]_+ - [M]_-$ in which both $[M]_+ \succeq 0$ and $[M]_- \succeq 0$. Let θ be the angle between $\mathbf{e}$ and $\hat{\mathbf{X}}y - w$. By Lemma 14 in [101], we have

$$\operatorname{tr}([M]_+) = \|\mathbf{e}\|\|\hat{\mathbf{X}}y - w\|(1 + \cos\theta), \quad \operatorname{tr}([M]_-) = \|\mathbf{e}\|\|\hat{\mathbf{X}}y - w\|(1 - \cos\theta).$$

Now, one can verify that $(U_1^*, U_2^*, W^*, G^*, \lambda^*, y^*)$ defined as

$$\begin{gathered} U_1^* = \frac{[M]_+}{\operatorname{tr}([M]_+)}, \quad U_2^* = \frac{[M]_-}{\operatorname{tr}([M]_+)}, \quad y^* = \frac{y}{\operatorname{tr}([M]_+)}, \\ W^* = \frac{W}{\operatorname{tr}([M]_+)}, \quad \lambda^* = \frac{\|y^*\|}{(2\epsilon + \kappa)\|\hat{X}\|_2}, \quad G^* = \frac{1}{\lambda^*} y^* y^{*T} \end{gathered}$$

forms a feasible solution to the dual problem (6.20) *whose objective value is equal to*

$$\frac{\operatorname{tr}([M]_-) + \langle \hat{\mathbf{X}}^\top \hat{\mathbf{X}}, W \rangle + (2\epsilon + \kappa)\operatorname{tr}(W) + (4\epsilon + 2\kappa)\|\hat{X}\|_2\|y\|}{\operatorname{tr}([M]_+)}.$$

Substituting (6.26), (6.28), (6.29), *and* (6.30) *into the above equation, we obtain*

$$\begin{aligned} \eta^*(\hat{X}) &\leq \frac{2\beta\gamma + 1 - \psi(\gamma) + ((2\epsilon + \kappa)\sqrt{r} + 2\sqrt{(2\epsilon + \kappa)(1 + \delta)}\|\hat{X}\|_2)/\|\mathbf{e}\|}{1 + \psi(\gamma)} \\ &\leq \frac{2\beta\gamma + 1 - \psi(\gamma)}{1 + \psi(\gamma)} + \frac{(2\epsilon + \kappa)\sqrt{r} + 2\sqrt{(2\epsilon + \kappa)(1 + \delta)}\|\hat{X}\|_2}{\|\mathbf{e}\|}. \end{aligned}$$

Choosing the best value of the parameter $\gamma \in [0, 1]$ to minimize the far right-side of the above inequality leads to

$$\frac{2\beta\gamma + 1 - \psi(\gamma)}{1 + \psi(\gamma)} \leq \eta_0(\hat{X}),$$

with

$$\eta_0(\hat{X}) := \begin{cases} \dfrac{1 - \sqrt{1 - \alpha^2}}{1 + \sqrt{1 - \alpha^2}}, & \text{if } \beta \geq \dfrac{\alpha}{1 + \sqrt{1 - \alpha^2}}, \\ \dfrac{\beta(\alpha - \beta)}{1 - \beta\alpha}, & \text{if } \beta \leq \dfrac{\alpha}{1 + \sqrt{1 - \alpha^2}}. \end{cases}$$

Here, α and β are defined in (6.25) *and* (6.27), *respectively. In the proof of Theorem 1.2 in [95], it is shown that*

$$\frac{1 - \eta_0(\hat{X})}{1 + \eta_0(\hat{X})} \geq \frac{1}{1 + \sqrt{r/r_{search}}}$$

for every $\hat{X}$ with $\hat{X}\hat{X}^\top \neq ZZ^\top$ and $\operatorname{rank}(Z) = r$. *Therefore,*

$$\eta_0(\hat{X}) \leq \frac{\sqrt{r/r_{search}}}{2 + \sqrt{r/r_{search}}},$$

which gives the upper bound (6.18).

Finally, we still need to deal with the case when $\mathcal{P}_\perp Z = 0$. In this case, we know that $\hat{\mathbf{X}}\hat{y} = \mathbf{e}$ with $\hat{y}$ defined in (6.21)*. Then, it is easy to check that $(U_1^*, U_2^*, W^*, G^*, \lambda^*, y^*)$ defined as*

$$U_1^* = \frac{\mathbf{e}\mathbf{e}^\top}{\|\mathbf{e}\|^2}, \quad U_2^* = 0, \quad y^* = \frac{\hat{y}}{2\|\mathbf{e}\|^2},$$
$$W^* = 0, \quad \lambda^* = \frac{\|y^*\|}{(2\epsilon + \kappa)\|\hat{X}\|_2}, \quad G^* = \frac{1}{\lambda^*} y^* y^{*T}$$

forms a feasible solution to the dual problem (6.20) *whose objective value is* $(4\epsilon + 2\kappa)\|\hat{X}\|_2\|y^*\|$*. By the inequality* (6.23)*, we have*

$$\eta^*(\hat{X}) \leq (4\epsilon + 2\kappa)\|\hat{X}\|_2\|y^*\| \leq \frac{(2\epsilon + \kappa)\|\hat{X}\|_2}{\sqrt{2}\sigma_{r_{search}}(\hat{X})\|\mathbf{e}\|} \leq \frac{\sqrt{(2\epsilon + \kappa)(1 + \delta)}\|\hat{X}\|_2}{\|\mathbf{e}\|}$$

Hence, the upper bound (6.18) *still holds in this case.*

Finally, Theorem 14 is a direct consequence of Lemma 24 and Lemma 25. Theorem 11 is a special case of Theorem 14 with $\kappa = 0$, and the global convergence in Theorem 15 is also a corollary of Theorem 14.

Proof 31 (Proof of Theorem 15) *Assume that $\|\mathbf{A}^\top w\| \leq \epsilon/2$ holds. Since*

$$\nabla^2 h_w(\hat{X}) = 2I_{r_{search}} \otimes \operatorname{mat}_S(\mathbf{A}^\top(\mathbf{A}\mathbf{e} + w)) + \hat{\mathbf{X}}^\top\mathbf{A}^\top\mathbf{A}\hat{\mathbf{X}} = \nabla^2 h_0(\hat{X}) + 2I_{r_{search}} \otimes \operatorname{mat}_S(\mathbf{A}^\top w),$$

and

$$|\lambda_{\max}(I_{r_{search}} \otimes \operatorname{mat}_S(\mathbf{A}^\top w))| \leq \|\operatorname{mat}_S(\mathbf{A}^\top w)\|_F \leq \|\mathbf{A}^\top w\| \leq \frac{\epsilon}{2}$$

as shown in the proof of Lemma 23, by the assumption it holds that

$$\begin{aligned} -\lambda_0 > \lambda_{\min}(\nabla^2 h_0(\hat{X})) &\geq \lambda_{\min}(\nabla^2 h_w(\hat{X})) - 2\lambda_{\max}(I_{r_{search}} \otimes \operatorname{mat}_S(\mathbf{A}^\top w)) \\ &\geq \lambda_{\min}(\nabla^2 h_w(\hat{X})) - \epsilon \end{aligned} \tag{6.31}$$

for every matrix $\hat{X} \in \mathbb{R}^{n \times r_{search}}$ with $\|\hat{X}\|_2 < D$. The perturbed gradient descent method in [31] will find a solution $\hat{X}$ satisfying (6.11) *with*

$$\tilde{\kappa} = \min\{\lambda_0 - \epsilon, D\epsilon\}$$

in $O(\mathrm{poly}(1/\epsilon))$ *number of iterations. The inequality* (6.31) *and the second condition in* (6.11) *together imply that* $\|\hat{X}\|_2 \geq D$*, and thus the conditions in* (6.8) *are automatically satisfied for* $\hat{X}$ *with* $\kappa = \epsilon$*, which gives the desired result after we apply Theorem 14 with the original* ϵ *in Theorem 14 replaced with* $\epsilon/2$ *and* κ *replaced with* ϵ*.*

The proof of Theorem 12 is similar to the above proof of Lemma 25 in the situation with $\kappa = 0$, and we will only emphasize the difference here.

Proof 32 (Proof of Theorem 12) *In the case when* $\hat{X} \neq 0$*, after constructing the feasible solution to the dual problem* (6.20)*, we have*

$$\frac{1-\delta}{1+\delta} \leq \eta^*(\hat{X}) \leq \frac{\mathrm{tr}([M]_-) + \langle \hat{\mathbf{X}}^\top \hat{\mathbf{X}}, W \rangle + 2\epsilon\,\mathrm{tr}(W) + 4\|\hat{X}\|_2 \epsilon \|y\|}{\mathrm{tr}([M]_+)}. \tag{6.32}$$

Note that in the rank-1 case, one can write $\sigma_{r_{search}}(\hat{X}) = \|\hat{X}\|_2$ *and*

$$\|y\| \leq \frac{\|\hat{y}\|}{\|\mathbf{e}\| \|\hat{\mathbf{X}}\hat{y}\|} \leq \frac{1}{\sqrt{2}\|\hat{X}\|_2 \|\mathbf{e}\|},$$

in which the last inequality is due to (6.23)*. Substituting* (6.26)*,* (6.28)*,* (6.29) *and the above inequality into* (6.32) *and choosing an appropriate* γ *as shown in the proof of Lemma 25, we obtain*

$$\begin{aligned} \frac{1-\delta}{1+\delta} \leq \eta^*(\hat{X}) &\leq \frac{2\beta\gamma + 1 - \psi(\gamma) + (2\epsilon + 2\sqrt{2}\epsilon)/\|\mathbf{e}\|}{1+\psi(\gamma)} \\ &\leq \frac{1}{3} + \frac{2\epsilon + 2\sqrt{2}\epsilon}{\|\mathbf{e}\|}, \end{aligned}$$

which implies inequality (6.4) *under the probabilistic event that* $\|\mathbf{A}^\top w\| \leq \epsilon$*.*

In the case when $\hat{X} = 0$*,* $(U_1^*, U_2^*, W^*, G^*, \lambda^*, y^*)$ *with*

$$\begin{aligned} U_1^* &= \frac{\mathbf{e}\mathbf{e}^\top}{\|\mathbf{e}\|^2}, \quad U_2^* = 0, \quad y^* = 0, \\ W^* &= \frac{ZZ^\top}{2\|\mathbf{e}\|^2}, \quad \lambda^* = 0, \quad G^* = 0 \end{aligned}$$

forms a feasible solution to the dual problem (6.20)*, which shows that*

$$\frac{1-\delta}{1+\delta} \leq \eta^*(\hat{X}) \leq \frac{\epsilon}{\|\mathbf{e}\|}.$$

The above inequality also implies inequality (6.4) *under the probabilistic event that* $\|\mathbf{A}^\top w\| \leq \epsilon$*.*

Now, we turn to the proof of the local guarantee in Theorem 13.

Proof 33 (Proof of Theorem 13) *Similar to the proof of Theorem 11, we assume that the probabilistic event $\|\mathbf{A}^\top w\| \leq \epsilon$ occurs and also break down the analysis into two cases. Consider the case when*

$$\sigma_{r_{search}}(\hat{X}) > \sqrt{\frac{\epsilon}{1+\delta}},$$

otherwise it is already handled by Lemma 24 that leads to the inequality (6.6)*. Here, we further relax the optimization problem* (6.15) *in Lemma 25 with $\kappa = 0$ by removing the constraint corresponding to the second-order optimality condition, which gives rise to the optimization problem*

$$\begin{aligned} \max_{\eta, \hat{\mathbf{H}}} \quad & \eta \\ \text{s.t.} \quad & \|\hat{\mathbf{X}}^\top \hat{\mathbf{H}} \mathbf{e}\| \leq 2\epsilon \|\hat{X}\|_2, \\ & \eta I_{n^2} \preceq \hat{\mathbf{H}} \preceq I_{n^2}. \end{aligned} \tag{6.33}$$

By denoting the optimal value of (6.33) *as $\eta_f^*(\hat{X})$, it holds that*

$$\eta_f^*(\hat{X}) \geq \eta^*(\hat{X}) \geq \frac{1-\delta}{1+\delta}. \tag{6.34}$$

Without loss of generality, we can assume that $\hat{X}$ is in the block form

$$\begin{bmatrix} X_1 \\ 0 \end{bmatrix}$$

with $X_1 \in \mathbb{R}^{r \times r}$ being invertible. Otherwise, there exists an orthogonal matrix $Q \in \mathbb{R}^{n \times n}$ such that $Q^\top \hat{X}$ satisfies this requirement. We can then replace $\hat{X}$ and M^* with $Q^\top \hat{X}$ and $Q^\top M^* Q$ respectively due to the invariance of (6.33) *under the transformation. Moreover, we select a matrix $Z \in \mathbb{R}^{n \times r_{search}}$ such that $ZZ^\top = M^*$ and Z is in the form*

$$\begin{bmatrix} Z_1^* & 0 \\ Z_2^* & 0 \end{bmatrix}$$

with $Z_1^* \in \mathbb{R}^{r \times r}$, $Z_2^* \in \mathbb{R}^{(n-r) \times r}$. Then, $\|Z_1^*(Z_2^*)^\top\|_F^2 \geq \lambda_r((Z_1^*)^\top Z_1^*)\|Z_2^*\|_F^2$, and

$$\begin{aligned} \lambda_r((Z_1^*)^\top Z_1^*) &\geq \lambda_r((Z_1^*)^\top Z_1^* + (Z_2^*)^\top Z_2^*) - \lambda_1((Z_2^*)^\top Z_2^*) \\ &\geq \lambda_r(Z^\top Z) - \|Z_2^*(Z_2^*)^\top\|_F \\ &\geq (1-\tau)\lambda_r(M^*) > 0, \end{aligned} \tag{6.35}$$

in which the last inequality is due to the assumption $0 < \tau < 1$ *and the second last inequality is due to*

$$\begin{aligned}\|Z_2^*(Z_2^*)^\top\|_F &\leq (\|Z_1^*(Z_1^*)^\top - X_1X_1^\top\|_F^2 + 2\|Z_1^*(Z_2^*)^\top\|_F^2 + \|Z_2^*(Z_2^*)^\top\|_F^2)^{1/2} \\ &= \|\hat{X}\hat{X}^\top - ZZ^\top\|_F \leq \tau\lambda_r(M^*).\end{aligned} \tag{6.36}$$

To prove the inequality (6.7)*, we need to bound* $\eta_f^*(\hat{X})$ *from above, which can be achieved by finding a feasible solution to the dual problem of* (6.33) *given below:*

$$\begin{aligned}\min_{U_1,U_2,G,\lambda,y} \quad & \operatorname{tr}(U_2) + 4\epsilon^2\|\hat{X}\|_2^2\lambda + \operatorname{tr}(G) \\ \text{s. t.} \quad & \operatorname{tr}(U_1) = 1, \\ & (\hat{\mathbf{X}}y)\mathbf{e}^\top + \mathbf{e}(\hat{\mathbf{X}}y)^\top = U_1 - U_2, \\ & \begin{bmatrix} G & -y \\ -y^\top & \lambda \end{bmatrix} \succeq 0, \\ & U_1 \succeq 0, \quad U_2 \succeq 0.\end{aligned} \tag{6.37}$$

If we choose $\hat{Y}$ *and* $\hat{y} = \operatorname{vec}(\hat{Y})$ *as* (6.21) *in the proof of Lemma 25, and let* θ *be the angle between* $\hat{\mathbf{X}}\hat{y}$ *and* $\mathbf{e}$*, then* (6.22) *implies that*

$$\begin{aligned}\sin^2\theta = \frac{\|\hat{\mathbf{X}}\hat{y} - \mathbf{e}\|^2}{\|\mathbf{e}\|^2} &= \frac{\|\mathcal{P}_\perp ZZ^\top\mathcal{P}_\perp\|_F^2}{\|\hat{X}\hat{X}^\top - ZZ^\top\|_F^2} \\ &= \frac{\|Z_2^*(Z_2^*)^\top\|_F^2}{\|Z_1^*(Z_1^*)^\top - X_1X_1^\top\|_F^2 + 2\|Z_1^*(Z_2^*)^\top\|_F^2 + \|Z_2^*(Z_2^*)^\top\|_F^2}.\end{aligned}$$

Following an argument similar to the one at the end of the proof of Lemma 7 in [93] and using (6.35) *and* (6.36)*, we can obtain*

$$\sin^2\theta \leq \frac{\tau}{2-\tau} \leq \tau. \tag{6.38}$$

Define

$$M = (\hat{\mathbf{X}}\hat{y})\mathbf{e}^\top + \mathbf{e}(\hat{\mathbf{X}}\hat{y})^\top,$$

and then decompose M *as* $M = [M]_+ - [M]_-$ *with* $[M]_+ \succeq 0$ *and* $[M]_- \succeq 0$*. Then, it is easy to verify that* $(U_1^*, U_2^*, G^*, \lambda^*, y^*)$ *defined as*

$$y^* = \frac{\hat{y}}{\operatorname{tr}([M]_+)}, \quad U_1^* = \frac{[M]_+}{\operatorname{tr}([M]_+)}, \quad U_2^* = \frac{[M]_-}{\operatorname{tr}([M]_+)},$$
$$G^* = \frac{y^*(y^*)^\top}{\lambda^*}, \quad \lambda^* = \frac{\|y^*\|}{2\epsilon\|\hat{X}\|_2}$$

forms a feasible solution to the dual problem (7.32) *with the objective value*

$$\frac{\operatorname{tr}([M]_-) + 4\epsilon\|\hat{X}\|_2\|\hat{y}\|}{\operatorname{tr}([M]_+)}. \tag{6.39}$$

Furthermore, it follows from the Wielandt–Hoffman theorem that

$$|\lambda_1(\hat{X}\hat{X}^\top) - \lambda_1(M^*)| \leq \|\hat{X}\hat{X}^\top - M^*\|_F \leq \tau\lambda_r(M^*).$$

Thus, using the above inequality, the inequality (6.23) *and the assumption on* $\sigma_{r_{search}}(\hat{X})$*, we have*

$$\frac{2\|\hat{X}\|_2\|\hat{y}\|}{\|\hat{\mathbf{X}}\hat{y}\|} \leq \frac{2\|\hat{X}\|_2}{\sqrt{2}\sigma_{r_{search}}(\hat{X})} \leq \sqrt{\frac{2(1+\delta)(\lambda_1(M^*) + \tau\lambda_r(M^*))}{\epsilon}} = C(\tau, M^*)\sqrt{\frac{1+\delta}{\epsilon}}. \tag{6.40}$$

Next, according to Lemma 14 of [101], one can write

$$\begin{aligned}\operatorname{tr}([M]_+) &= \|\hat{\mathbf{X}}\hat{y}\|\|\mathbf{e}\|(1+\cos\theta),\\ \operatorname{tr}([M]_-) &= \|\hat{\mathbf{X}}\hat{y}\|\|\mathbf{e}\|(1-\cos\theta).\end{aligned}$$

Substituting the above two equations and (7.35) *into the dual objective value* (7.33)*, one can obtain*

$$\eta_f^*(\hat{X}) \leq \frac{1 - \cos\theta + 2\sqrt{\epsilon(1+\delta)}C(\tau, M^*)/\|\mathbf{e}\|}{1+\cos\theta},$$

which together with (7.31) *implies that*

$$\|\mathbf{e}\| \leq \sqrt{\epsilon}(1+\delta)^{3/2}C(\tau, M^*)(\cos\theta - \delta)^{-1}.$$

The inequality (6.7) *can then be proved by combining the above inequality and* (6.38).

Chapter 7

Noisy General Low-Rank Recovery

7.1 Background and Related Work

The investigation of noise influence on our standard matrix sensing objective (1.3) performed in Chapter 6 inspires us to further generalize our results to a wider range of low-rank matrix problems. In this chapter, we focus on the problem of general noisy matrix optimization:

$$\min_{M\in\mathbb{R}^{n\times n}} f(M,w) \text{ s.t. } \operatorname{rank}(M)\leq r, M\succeq 0, \tag{7.1}$$

where the objective f takes in two input variables: a low-rank, positive semidefinite matrix $M\in\mathbb{R}^{n\times n}$ and a random variable $w\in\mathbb{R}^m$ that represents some corruption to the objective function. The noise can come from any arbitrary distribution as long as it has a finite variance. We denote the maximum rank of the variable M to be r. We optimize (7.1) only with respect to the first variable M, while w is assumed to be hidden to the user. The randomness of the parameter w comes from the stage prior to solving (7.1), which accounts for uncertainty in the model/data or external factors. Therefore, when the non-convex low-rank optimization is performed, w will not change anymore even though it is unknown to the user. Let M^* be a rank-r matrix that minimizes the function $f(M,0)$ subject to the constraints in (7.1). This problem has a wide range of applications, the most notable ones being matrix sensing [72], matrix completion [15], and robust PCA [12]. This formulation also has extensive applications in recommender systems [42], motion detection [23, 3], phase synchronization/retrieval [75, 8, 74], and power system estimation [104]. The matrix M^* is called the ground truth solution since the objective function is set up to be nonnegative and that $f(M^*,0)=0$ for most of the above-mentioned applications. The goal is to find the matrix closest to M^* in terms of Frobenius norm under the rank constraint. However, the influence of noise is not well studied in the literature due to the complications it may bring.

The major innovation of this work is the analysis of the effect of noise, where the objective function is subject to random corruption that is unknown to the user. This formulation is important yet oftentimes glossed over due to its challenging mathematical analysis, partly due to the sophisticated relationship between each globally optimal solution M and the vector w. For instance, consider the canonical example of the matrix sensing problem (6.1) given in the previous chapter, where $\tilde{b} = \mathcal{A}(M^*) - w$ represents perfect measurements on some ground truth M^* plus some noise w. The user only observes $\tilde{b}$ and has no access to noiseless measurements, which means that the matrix M^* of interest is the global minimum of (6.1) only when $w = 0$. When $w \neq 0$, the global minimum of (6.1) would likely differ from M^*. In this case, it is desirable to study whether local search algorithms can converge to a point that is close to M^* with high probability. Other applications such as matrix completion and robust PCA also suffer from the same conundrum since they all aim to align a given matrix to some partially observed matrix that is corrupted by unknown noise. In real-life problems, the corruptions induced by noise cannot be ignored or circumvented because they usually come from physical sources. For instance, in the power grid state estimation problem, which can be formulated as matrix sensing [36], measurements come from physical devices and the noise can be originated from mechanical failures, bad weather, and even cyber-attacks.

Due to the existence of a rank constraint, the optimization problem (7.1) is non-convex. Thus, local search algorithms can potentially converge to poor local minimizers, defeating the purpose of solving (7.1). Although (7.1) may be solved via convex relaxations for different classes of $f(\cdot, \cdot)$ to overcome the non-convexity issue when $f(\cdot, 0)$ is quadratic [15, 72, 16], the computational challenge associated with solving semidefinite programming problems is prohibitive for large-scale problems. This has inspired many papers to solve (7.1) via the Burer-Monteiro factorization [10] by factoring M into $XX^\top$, where $X \in \mathbb{R}^{n \times r}$, since M is positive semidefinite and has rank at most r. By doing so, one can convert the constrained optimization (7.1) into an unconstrained problem. Specifically, we solve the following problem instead of (7.1):

$$\min_{X \in \mathbb{R}^{n \times r}} f(XX^\top, w) \tag{7.2}$$

The main issue with (7.2) is that it is still a non-convex problem, despite being more scalable and easier to deal with computationally. To address this issue, a popular line of research in the literature is to study the optimization landscape of (7.2). Namely, the goal is to find the distance between the furthest local minimum and the global minimum, in addition to studying the convergence rate of local search methods in terms of the geometry of the optimization landscape. *Note here we assume we have access to the true rank r.*

7.1.1 Related Works

We first discuss the line of work that focuses on certifying the in-existence of spurious local minima in the noiseless setting (a local minimum that is not a global minimum is called *spurious*). [5] analyzes the absence of spurious local minima under the RIP condition for the matrix sensing problem, or in other words when $f(\cdot, w)$ is quadratic. This study states that $\delta \leq 1/5$ is a sufficient condition. [105, 46] investigate arbitrary objective functions under the RIP constant $\delta \leq 1/5$. The series of work [101, 93] show that the bound $\delta < 1/2$ is a sharp bound for guaranteeing the absence of spurious local minima in the case when the objective function is quadratic. The state-of-the-art result for general objective functions is proved in [7]'s paper, which states that $\delta < 1/2$ is also sufficient for the absence of spurious local minima.

In the noisy case, [102] proves that all local minima are close to the ground truth when $\delta \leq 1/35$ for a general objective, which is an extremely strong assumption on δ. Furthermore, [102] requires the RIP condition to be satisfied for the noisy problem rather than its noiseless counterpart, which is impossible to verify beforehand due to the unknown noise. For specific objective functions in the form of (6.1), Chapter 6 shows that $\delta < 1/2$ is sufficient and necessary for the absence of spurious local minima, even when w is sampled from an arbitrary finite-variance family. The major difference between Chapter 6 and this chapter is that we focus on a general objective, while Chapter 6 only focuses on a quadratic objective (the matrix sensing objective).

In terms of the convergence of local search methods, [82] proves that the gradient descent algorithm applied to (7.2) converges linearly when the initialization is good, given that $\delta < 1/7$. Similarly to [102], this RIP bound is given with respect to the noisy problem rather than its noiseless version, which is an undesirable feature. For a general noiseless objective (in the case when $w = 0$), [7] proves that there exists a region around M^* in which linear convergence can be established. On the other hand, [7] also proves that a RIP constant of $\delta < 1/2$ is sufficient for the global establishment of the strict saddle property for a general noiseless objective function. As noted in [31], the strict saddle property can lead to a polynomial convergence to a global optimum with a random initialization. The exact definition of the strict saddle property can be found in [24], and it basically states that all approximate local optima must be close to the global optima. In the noisy setting, Chapter 6 demonstrates that $\delta < 1/2$ is necessary and sufficient to the establishment for the strict saddle property for quadratic objective functions.

To highlight our improvements over the existing results, Table 7.1 lists some state-of-the-art comparable works to showcase the strength of the guarantees provided in this paper. Note that when we denote the objective function as "General Noisy", it means that the function is in the form of (7.2) and satisfies the RIP property. We further denote the objective function of (6.1) as "Quadratic Noisy", which is also known as the matrix sensing problem. In particular, according to [14], $\mathcal{O}(1/\delta^2)$ number of random Gaussian measurements are required to ensure δ-RIP$_{2r}$, so a RIP

Table 7.1: Comparison between our result and the prior literature.

Paper	Objective function	Local Min	Strict Saddle	Convergence Rate
[102]	General Noisy	$\delta < 1/35$	N/A	N/A
[82]	General Noisy	N/A	N/A	Linear when $\delta < 1/7$
Chapter 6	Quadratic Noisy	$\delta \leq 1/2$	$\delta < 1/2$	N/A
Ours	General Noisy	$\delta < 1/3$	$\delta < 1/3$	Linear with general δ

constant of 1/3 vs 1/35 introduces a difference in sample number requirement of around $35^2/3^2 \approx 100$ times. Also since [102] requires δ-RIP$_{6r}$, even more measurements are needed.

7.1.2 Assumptions on the objective function

The assumptions stated in this section serve as the underpinnings of all the theorems in this paper, and they mainly require that the objective function be smooth with respect to both the decision variable X and the noise w. To clarify, these assumptions do not pose any restriction on w, and this parameter can come from any probability distribution.

Assumption 1. The objective function $f(\cdot,\cdot)$ is twice continuously differentiable with respect to its first argument M.

Assumption 2. The noiseless objective function $f(\cdot,0)$ satisfies the δ-RIP$_{2r,2r}$ property for some constant $\delta \in [0,1)$.

Assumption 3. The noise w has a finite influence on the gradient and Hessian of the objective function in the sense that there exist two constants $\zeta_1 \geq 0$ and $\zeta_2 \geq 0$ such that

$$|\langle \nabla_M f(M,w) - \nabla_M f(M,0), K\rangle| \leq \zeta_1 \|w\|_2 \|K\|_F, \tag{7.3}$$

$$|[\nabla^2_M f(M,w) - \nabla^2_M f(M,0)](K,L)| \leq \zeta_2 \|w\|_2 \|K\|_F \|L\|_F \tag{7.4}$$

for all matrices $M, K, L \in \mathbb{R}^{n\times n}$ with $\operatorname{rank}(M), \operatorname{rank}(K), \operatorname{rank}(L) \leq 2r$.

As an example, for the standard matrix sensing problem (6.1) with the sensing matrix $\mathcal{A}$, if $\mathcal{A}$ satisfies the RIP property, then all of these assumptions hold with $\zeta_1 = \|\mathcal{A}\|_2$ and $\zeta_2 = 0$. The 1-bit matrix completion problem is also an example that satisfies the above assumptions which will be elaborated in Section 7.4.2. Note that although in our problem statements we assume M to be symmetric and positive semidefinite, our framework can also be adapted to deal with non-symmetric and non-square matrix M. A more detailed discussion is provided in Section 2.2.

7.2 Landscape of General Noisy Problems

7.2.1 When $\delta < 1/3$

When the RIP constant δ is smaller than 1/3, we show that all local minima (or second-order critical points) of (7.2) are close to the ground truth solution M^*. The proximity to the ground truth is parametrized by the noise intensity defined as $q := \|w\|_2$. When $q = 0$, our result (Theorem 16) recovers the results previously proved in [28, 94].

Theorem 16 *Assume that the objective function of* (7.2) *satisfies Assumptions 1-3 and that* $f(M, 0)$ *satisfies the RIP property with some* δ*-RIP*$_{2r,2r}$ *constant such that* $\delta < 1/3$. *For every* $\epsilon \in [0, \frac{1/3-\delta}{\zeta_2})$, *with probability at least* $\mathbb{P}(\|w\|_2 \leq \epsilon)$, *every local minimizer* $\hat{X}$ *of* (7.2) *satisfies:*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{2\zeta_1\epsilon}{1 - 3(\delta + \zeta_2\epsilon)}. \tag{7.5}$$

This is a powerful theorem stating that as long as $\delta < 1/3$, all local minima are close to the ground truth solution, regardless of the family from which w is sampled. Previously, the problem needed to satisfy $\delta < 1/35$ for similar results to hold. Furthermore, unlike [6], we achieved this result without the BDP assumption or requiring $r = 1$. The upper bound in (7.5) is a function of ϵ and δ. The bound becomes loser as ϵ and δ increase. Note that ζ_1 and ζ_2 affect ϵ in a linear way and therefore obtaining non-conservative constants ζ_1 and ζ_2 is beneficial.

Our result implies that for a general objective function, geometric uniformity, captured by the RIP property, can guarantee a benign optimization landscape even when δ is non-trivially larger than 0. However, this comes with a caveat. In particular, if $\zeta_2 \neq 0$, meaning that the Hessian is affected by the existence of noise, then there is a hard "contribution floor" for the noise reflected by the inequality $\|w\|_2 \leq \frac{1/3-\delta}{\zeta_2}$. If the noise intensity goes beyond this hard limit, no high-probability guarantees can be made in terms of the locations of the local minima. This is expected because if ζ_2 is large, it means that the RIP property satisfied for the noiseless problem cannot enforce any desirable property on the highly noisy problem and the benign optimization landscape is unlikely to hold.

The proof of Theorem 16 follows from the characterization of the r-th singular value of an arbitrary local minimizer $\hat{X}$. Previous results in the literature successfully upper-bounded the r-th singular value of X that are far from the ground truth, which leads to the establishment of a significant escape direction based on its Hessian. The major innovation in the proof of Theorem 16 is based on the observation that for every local minimizer $\hat{X}$, its r-th singular value can also be lower-bounded in terms of the smallest eigenvalue of the gradient at $\hat{X}$, and the RIP constant. Then we adopt some existing techniques to also upper-bound the r-th singular value of $\hat{X}$ to contrast it with the lower-bound. By doing so, we derive necessary conditions on the value of $\|\hat{X}\hat{X}^\top - M^*\|$, since the upper-bounds are carefully crafted to include this term. We believe that this new method of lower-bounding the r-th singular value of $\hat{X}$ could open up a new range of possible techniques for analyzing low-rank optimization problems, since it provides important complementary information on $\hat{X}$. The full proof is lengthy and deferred to Appendix 7.A.1.

7.2.2 When $\delta \geq 1/3$

Although Theorem 16 is powerful in the case of $\delta < 1/3$, it does not provide any guarantee when $\delta \geq 1/3$, especially given the fact that δ is intrinsic to the sensing matrices, which are impossible to change. This is where a local version of the guarantee comes in handy. We only consider the optimization landscape in a region around the ground truth and show that local minimizers are all very close to M^*.

Theorem 17 *Assume that the objective function of* (7.2) *satisfies assumptions 1-3 with $f(M, 0)$ satisfying the δ-RIP$_{2r,2r}$ property for a constant $\delta \in [0, 1)$. Consider an arbitrary number $\tau \in (0, 1-\delta^2)$. Every local minimizer $\hat{X} \in \mathbb{R}^{n\times r}$ of* (7.2) *satisfying:*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \tau\lambda_r(M^*), \tag{7.6}$$

will also satisfy the following inequality with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$:

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{\epsilon(1+\delta+\zeta_2\epsilon)\zeta_1 C(\tau, M^*)}{\sqrt{1-\tau} - \zeta_2\epsilon - \delta} \tag{7.7}$$

for all $\epsilon < \frac{\sqrt{1-\tau}-\delta}{\zeta_2}$, where

$$C(\tau, M^*) = \sqrt{\frac{2(\lambda_1(M^*) + \tau\lambda_r(M^*))}{(1-\tau)\lambda_r(M^*)}}.$$

The upper bounds in (7.6) and (7.7) define an outer ball and an inner ball centered at the ground truth M^*. Theorem 17 asserts the absence of local minima in the ring between the two balls. As ϵ goes to 0, Theorem 17 states that no spurious local minima exists when $\|\hat{X}\hat{X}^\top - M^*\|_F \leq (1-\delta^2)\lambda_r(M^*)$. Therefore, this is a direct generalization of the results in [6], which holds only for noiseless objectives. This local theorem allows for the analysis of highly non-convex objectives associated with δ close to 1. In particular, Theorem 17 states that even for highly non-convex objectives, the optimization landscape is benign in the vicinity of M^*. This means that if a good initial point is selected, local search algorithms can solve this highly non-convex problem and find a satisfactory approximate solution.

The breakthrough of the proof of this theorem relies on the establishment of Lemma 29, which states that for every local minimizer $\hat{X}$ of the noisy problem (7.2), there is a pseudo sensing matrix $\mathbf{H}$ such that $\hat{X}$ is an approximate local minimizer of a matrix sensing problem with the sensing operator $\mathbf{H}$. This serves as the basis of the ensuing proof techniques, which follow the idea of certifying the in-existence of spurious local minima, inspired by [95]. A detailed proof can be found in Appendix 7.A.2.

7.3 Convergence of General Noisy Problems

7.3.1 Linear Convergence with good initialization

To establish linear convergence for the noisy problem (7.2), an additional assumption is required:

Assumption 4. There exists a constant ρ such that the gradient of the function $f(\cdot, w)$ with respect to the first argument M is $\rho-$restricted Lipschitz continuous, meaning that:

$$\|\nabla_M f(M, w) - \nabla_M f(M', w)\|_F \leq \rho \|M - M'\|_F$$

for all matrices $M, M' \in \mathbb{R}^{n \times n}$ with $\operatorname{rank}(M) \leq r$ and $\operatorname{rank}(M') \leq r$.

Assumption 4 is critical for the convergence of local search algorithms since otherwise we cannot choose a step size small enough to avoid the constant overshoot of the algorithm. For a standard matrix sensing problem, $\nabla_M f(M, w) = \mathbf{A}\mathbf{A}^\top \operatorname{vec}(M) + \mathbf{A}^\top w$, hence satisfying Assumption 4 with $\rho = \sigma_{\max}(\mathbf{A}\mathbf{A}^\top)$.

We now present our main result in this section, which states that if the initialization is close enough to M^w, then the gradient descent algorithm will reach M^w or a low-rank projection of M^w at a linear rate. Here, M^w is defined to be the unique global minimum of (7.1) without the rank constraint. Since (7.1) is a strongly convex problem without the rank constraint, M^w always exists and is unique. Given Theorems 16 and 17, we can in turn guarantee that M^w is close to M^*, showing that the gradient descent algorithm reaches a neighborhood of M^* in a satisfactory rate.

Theorem 18 *The vanilla gradient descent method applied to* (7.2) *under Assumptions 1-4 converges to $\mathcal{P}_r(M^w)$, the best rank-r approximation of M^w, linearly up to* *a difference D_r if the initial point X_0 satisfies:*

$$\|X_0X_0^\top - M^w\|_F < C_w^2(1-\delta-\zeta_2\epsilon) - C_w\sqrt{\frac{1-\delta-\zeta_2\epsilon}{1+\delta+\zeta_2\epsilon}}D_r, \tag{7.8}$$

meaning that vanilla gradient descent will reach a point $\tilde{M}$ linearly with $\|\tilde{M} - \mathcal{P}_r(M^w)\|_F \geq D_r$, where

$$D_r = \|M^w - \mathcal{P}_r(M^w)\|_F, \quad C_w = \sqrt{2(\sqrt{2}-1)\sigma_r(M^w)}.$$

The linear convergence is also contingent on the fixed step size η satisfying:

$$\eta \leq \left(12\rho r^{(1/2)}\left(C\sqrt{(1-(\delta+\zeta_2\epsilon)^2} + \|M^w\|_F\right)\right)^{-1}, \tag{7.9}$$

for all $\epsilon < \frac{1-\delta}{\zeta_2}$ with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$, where $C = 2(\sqrt{2}-1)$.

The main challenge stemming from the introduction of noise is that the unconstrained global minimum of (7.1) may not necessarily be of rank-r (the rank of M^*) anymore. Therefore, since the Monteiro-Burer approach (7.2) can only search over matrices of rank at most r, we can only guarantee the convergence of any algorithm with respect to a rank-r matrix, which for our purpose we chose to be $\mathcal{P}_r(M^w)$. Thus, the radius of linear convergence depends on D_r, a constant quantifying how close M^w is to a rank-r matrix. In the special case that M^w is of rank at most r, D_r becomes 0 and our Theorem can be simplified. We summarize this special case via the following assumption:

Assumption 5. The objective function $f(\cdot, w)$ of (7.1) has a first-order critical point M^w for every w such that it is symmetric, positive semidefinite, and $\text{rank}(M^w) \leq r$.

Assumption 5 may not hold in general, but for specific problems, such as (6.1), this assumption is satisfied if the set $\{\mathbf{A}\,\text{vec}(N - M^*) \mid \text{rank}(N) \leq r\}$ spans $\mathbb{R}^m$. This is highly likely since $m \ll n^2$. According to Proposition 1 in [105], if Assumption 5 is met, M^w is the global minimum of (7.1). With this assumption, we can now introduce a useful Corollary:

Corollary 2 *The vanilla gradient descent method applied to* (7.2) *under Assumptions 1-5 converges to M^w linearly if the initial point X_0 satisfies:*

$$\|X_0X_0^\top - M^w\|_F < 2(\sqrt{2}-1)(1-\delta-\zeta_2\epsilon)\sigma_r(M^w), \tag{7.10}$$

with fixed step size η satisfying:

$$\eta \leq \left(12\rho r^{(1/2)}\left(C\sqrt{(1-(\delta+\zeta_2\epsilon)^2} + \|M^w\|_F\right)\right)^{-1}, \tag{7.11}$$

for all $\epsilon < \frac{1-\delta}{\zeta_2}$ with probability at least $\mathbb{P}(\|w\|_2 \leq \epsilon)$, where $C = 2(\sqrt{2}-1)$.

Prior to this theorem, it was possible to establish a linear convergence using the existing literature only when Assumption 5 holds and $\delta < 1/7$. Now, Theorem 18 allows for having an arbitrary δ, and generalized the guarantee to cases where Assumption 5 does not hold. Theorem 18 further implies that even starting from an arbitrary initial point, the gradient descent algorithm has a linear convergence in the final phase, given that the step size is small enough and that the noise intensity is not high. This further implies that if a linear convergence is not observed, the user could decrease the step size until a linear convergence is established. This is confirmed empirically in Section 7.4.2.

Theorem 18 is inspired by the observation that since we only search on a low-rank manifold, we may never really reach M^w (even in the asymptotic regime), thus by constraining the search space away from M^w, linear convergence can be established. The full proof is deferred to Appendix 7.A.3.

7.3.2 Strict Saddle Property

When $\delta < 1/3$, the noisy problem (7.2) exhibits the strict saddle property, meaning that all approximate second-order critical points are close to the global optimum of the optimization problem with high probability:

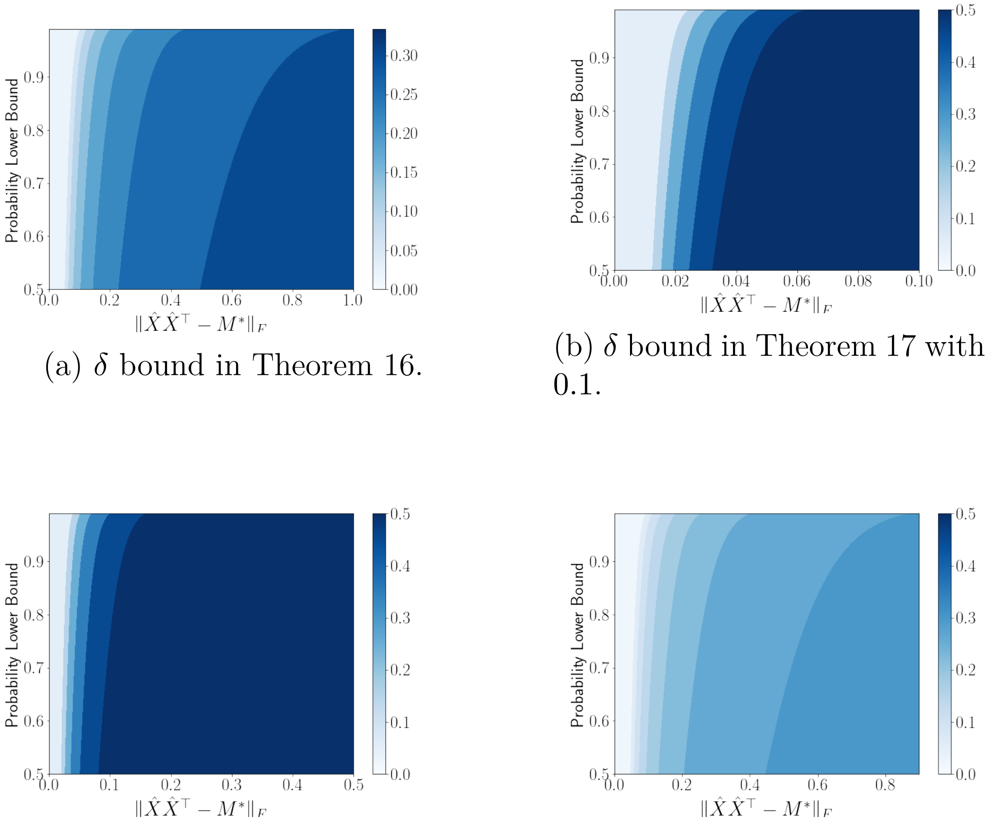
Theorem 19 *Suppose that the objective function of* (7.2) *satisfies assumptions 1-3 with a δ-RIP$_{2r,2r}$ constant of $\delta < 1/3$ in the noiseless case. Consider the ground truth solution M^* which is of rank r. For a given constant $\alpha > 0$, there exists a finite constant $\xi > 0$ such that at least one of the three following conditions holds for any $X \in \mathbb{R}^{n\times r}$:*

$$\begin{aligned}&\operatorname{dist}(X, M^*) \leq \alpha,\ \ \|\nabla_X h(X,w)\|_F \geq \xi,\\ &\lambda_{\min}(\nabla_X^2 h(X,w)) \leq -2\xi,\end{aligned}$$

with probability at least $\mathbb{P}(\|w\|_2 \leq \frac{1/3-\delta}{\zeta_2+2\zeta_\alpha/3})$, *where* $\zeta_\alpha := \zeta_1/(\sqrt{2(\sqrt{2}-1)}(\sigma_r(M^*))^{1/2}\alpha)$.

The significance of the establishment of the strict saddle property is that one can find an approximate local minimum in polynomial time. The perturbed gradient descent algorithm presented in [31] serves as one of the algorithms achieving this goal. Coupled with Theorem 18, it means that we could reach M^w with an arbitrary accuracy in polynomial time via a random initialization, which is also known to be close to M^* according to Theorems 16 and 17.

The proof of this theorem is similar to that of Theorem 7 of [94], and we highlight the key differences in Appendix 7.A.4 to illustrate how Theorem 19 can be proved.



(a) δ bound in Theorem 16.

(b) δ bound in Theorem 17 with $\tau = 0.1$.

(c) δ bound in Theorem 17 with $\tau = 0.5$.

(d) δ bound in Theorem 17 with $\tau = 0.9$.

Figure 7.1: Comparison of the maximum RIP constants δ allowed by Theorem 16 and Theorem 17 to guarantee a given bound on the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ for an arbitrary local minimizer $\hat{X}$ satisfying (7.6) with a given probability.

7.4 Numerical Illustration

In this section, we provide a concrete example to the results derived above[1]. We empirically study the proximity of an arbitrary local minimizer $\hat{X}$ of (7.2) to its ground truth solution in terms of $\|\hat{X}\hat{X}^\top - M^*\|_F$, and analyze the effect of the step size on the convergence rate.

Assume that $w \in \mathbb{R}^m$ is a $0.05/\sqrt{m}$-sub-Gaussian vector. According to Lemma 1 in [32], this choice of w satisfies:

$$1 - 2\mathrm{e}^{-\frac{\epsilon^2}{16m\sigma^2}} \leq \mathbb{P}(\|w\|_2 \leq \epsilon).$$

[1] Code used to produce the results in this section can be found here: https://github.com/anonpapersbm/Noisy-Low-rank-Matrix-Optimization

with $\sigma = 0.05$. We refer to the RHS of the above equation as the probability lower-bound since it says that the event $\|w\|_2 \le \epsilon$ will happen with probability at least that number.

7.4.1 Quality of Local Minima

We consider the problem of 1-bit Matrix Completion, which is a low-rank matrix optimization problem that naturally arises in recommendation systems with binary inputs [20, 26].

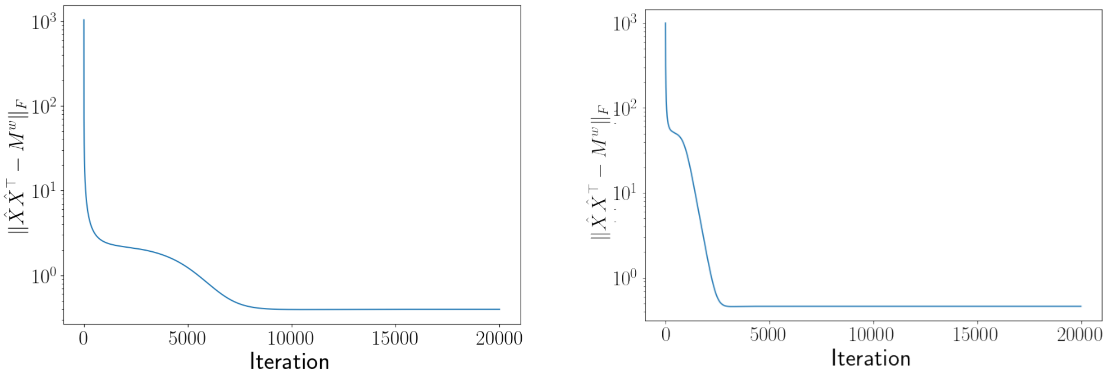
The objective of this 1-bit Matrix Completion problem is:

$$f(M, w) = -\sum_{i=1}^{n}\sum_{j=1}^{n}((y_{ij} + w_{ij})M_{ij} - \log(1 + \exp(M_{ij}))) \tag{7.12}$$

where M_{ij} is the $(i, j)^{\text{th}}$ component of M and $y_{ij} \in [0, 1]$ is a percentage-wise observation of M_{ij}. Since y_{ij} are empirical observations, they could very much be subject to random corruptions, which we explicitly represent by $w \in \mathbb{R}^m$, with $m = n^2$. It is straightforward to verify that (7.12) satisfies the assumptions outlined in Section 7.1.2, with $\zeta_1 = 1$ and $\zeta_2 = 0$.

The work [6] shows that for (7.12), the function $\gamma f(M, 0)$ exhibits the δ-RIP$_{2r,2r}$ property for some constant γ over the neighborhood $\|\hat{X}\hat{X}^\top - M^*\|_F \le R$ for a small R. We choose M^* such that $\lambda_r(M^*) = R$. Therefore, we can use the framework proposed in this paper to analyze the quality of the local minima of (7.12) under random perturbation.

In Figure 7.1, we numerically demonstrate and compare the bounds given in Theorem 16 and Theorem 17, for the parameters $n = 40$, and $r = 5$. We assume w comes from the sub-Gaussian distribution described above with $\sigma = 0.05$. The x-axis shows the maximum distance between an arbitrary local minimum $\hat{X}$ and the ground truth, and unit for the x-axis is $\lambda_r(M^*)$. The y-axis delineates the probability lower bound, which describes a lower-bound on the probability that the event will happen. The contour plot itself shows the maximum δ that is necessary to guarantee $\hat{X}$ to be in the range of $\|\hat{X}\hat{X}^\top - M^*\|_F \le \xi$ with ξ specified on the x-axis, with probability greater than the value specified on the y-axis. Figure 7.1 shows that as τ becomes smaller, for the same set of (x, y) values, the necessary value of δ becomes larger. This means that if the prior information on $\hat{X}$ is strong, meaning that it is known to lie within a neighborhood of the ground truth, then the local minima are tightly centered around the correct solution with a high probability. Moreover, the global bound is generally looser than that of the local version when τ is small, because it only applies to cases when $\delta < 1/3$, but when τ is large, the global bound could be better even with the same δ, as evident when comparing subfigures (a) and (d) in Figure 7.1.



(a) Convergence rate when step size is 0.001.

(b) Convergence rate when step size is 0.0002.

Figure 7.2: The distance to M^w versus iterations for gradient descent with random initialization.

Readers can refer to Appendix 7.B for more plots regarding the interplay of ζ_1, ζ_2 and σ values in Theorem 16. These values may represent a wide range of different objectives(each objective is characterized by its ζ_1 and ζ_2 values) and different noise patterns (characterized by σ value since many known distributions are sub-Gaussian).

7.4.2 Convergence Rate

In this section, we demonstrate the convergence rate of the vanilla gradient descent algorithm applied to an instance of (6.1) satisfying Assumptions 1-5 with $n = 40$, $m = 190$, $r = 5$. The matrix $\mathbf{A}$ used here makes the objective function satisfy 0.42-RIP$_{2r,2r}$. We also assume that $\lambda_1(M^*) = 1.5$ and $\lambda_r(M^*) = 1$. Note that (6.1) meets our assumptions with $\rho = \|\mathbf{A}\|_2^2, \zeta_1 = \|\mathbf{A}\|_2$, and $\zeta_2 = 0$. We aim to show how the step size affects the convergence rate, and corroborate the theoretical results in Theorem 18. Note since Assumption 5 is satisfied, the algorithm will converge to M^w directly.

In Figure 7.2, we choose two different step sizes, namely 0.001 and 0.0002, and start from random initialization. It can be observed that in the case of the larger step size, there is a region of plateauing in which the gradient descent algorithm makes little progress, while the smaller step size exhibits a linear convergence around iterations 500-2500 even after the initial phase of a fast descent. This result is in accordance with Corollary 2, which states that for a small enough step size, the gradient descent algorithm will achieve a linear convergence in a neighborhood of the global minimum.

7.5 Summary

In this chapter, we proposed a unified, yet general framework to analyze the global and local optimization landscapes of a class of noisy low-rank matrix optimization problems. We showed that regardless of the distribution from which the random noise is sampled, if the noiseless objective satisfies RIP, then there are mathematical guarantees on the locations of local minima and the convergence rate. This means that even for general objectives, geometric uniformity can compensate for random corruption. The results here significantly extends the existing results in the literature on this general problem, and offers new techniques and insights that can be used to study other noisy low-rank optimization problems.

7.A Missing Details of Chapter 7

Before diving into the proofs, we further impose some technical assumptions on our problem without loss of generality.

Assumption 1 *Assume that $\nabla_M f(M,w)$ is symmetric for every $M \in \mathbb{R}^{n\times n}$,*

This assumption always holds since otherwise we could simply optimize for $(f(M,w)+f(M^\top,w))/2$ instead. The parameter q is used to represent $\|w\|_2$ in the following proofs for the sake of notational simplicity.

We also make the following standard assumption:

Assumption 2 *The objective function $f(\cdot,0)$ of* (7.1) *has a first-order critical point M^* such that it is symmetric, positive semi-definite, and* $\operatorname{rank}(M^*) \leq r$.

This assumptions states that the objective in (7.1) can indeed recover the ground truth low rank matrix M^* when solved to global optimality. Otherwise solving for (7.1) under low rank constraint will be meaningless.

Note that

$$\nabla_M f(M^*,0) = 0 \tag{7.13}$$

is a consequence of Assumption 2 due to Proposition 1 in [105]. This proposition also implies that M^* is the unique global minimum of (7.1).

7.A.1 Proofs of Section 7.2.1

Lemma 28 *If $\hat{X}$ is a local minimum of* (7.2) *with $\hat{M} = \hat{X}\hat{X}^\top$, then*

$$\lambda_r^2(\hat{M}) \geq \frac{G^2}{(1+\delta+\zeta_2 q)^2} \tag{7.14}$$

where $G = -\lambda_{\min}(\nabla_M f(\hat{M},w))$.

Proof 34 (Proof of Lemma 28) *First consider the case where* $\operatorname{rank}(\hat{M}) = r$. *Under this assumption, consider the singular value decomposition (SVD) of $\hat{M}$:*

$$\hat{M} = \sum_{i=1}^{r} \sigma_i u_i u_i^\top,$$

where σ_i's are eigenvalues and u_i's are unit eigenvectors. Let u_G be a unit eigenvector of $\nabla f(\hat{M},w)$ such that $u_G^\top \nabla f(\hat{M},w) u_G = -G$. Furthermore, for a constant $p \in [0,1]$, define:

$$M_p = \sum_{i=1}^{r-1} \sigma_i u_i u_i^\top + \sigma_r (p u_G + \sqrt{1-p^2} u_r)(p u_G + \sqrt{1-p^2} u_r)^\top.$$

One can write:

$$\begin{aligned}\langle \nabla_M f(\hat{M}, w), M_p - \hat{M}\rangle &= \langle \nabla_M f(\hat{M}, w), \sigma_r p^2 u_G u_G^\top\rangle \\ &= -Gp^2\sigma_r.\end{aligned}$$

since $\nabla f(\hat{M}, w)u_i = \nabla f(\hat{M}, w)^\top u_i = 0 \ \forall i \in \{1, \dots, r\}$. *This is because* $\hat{X}$ *is a local minimum, and* (2.5) *is a necessary condition according to Lemma 4. We could choose a SVD of* $\hat{M}$ *such that:*

$$\hat{X} = \begin{bmatrix}\sigma_1^{1/2}u_1 & \sigma_2^{1/2}u_2 & \dots & \sigma_r^{1/2}u_r\end{bmatrix}.$$

Now, we expand the term $\|M_p - M\|_F^2$*:*

$$\begin{aligned}\|M_p - \hat{M}\|_F^2 =&\sigma_r^2 \operatorname{tr}\left((p^2 u_G u_G^\top + p\sqrt{1-p^2}u_G u_r^\top + p\sqrt{1-p^2}u_r u_G^\top - p^2 u_r u_r^\top)^2\right) \\ =&\sigma_r^2(p^4 + p^2(1-p^2) + p^2(1-p^2) + p^4) \\ =&2\sigma_r^2 p^2.\end{aligned}$$

where the second equality follows from the fact that $u_G^\top u_i = 0 \quad \forall i \in \{1, \dots, r\}$. *This is due to the fact that*

$$u_G^\top u_i = \left(\frac{-1}{G}\nabla_M f(\hat{M}, w)u_G\right)^\top u_i = 0.$$

This means that $\langle \nabla_M f(\hat{M}, w), M_p - \hat{M}\rangle = -\frac{G}{2\sigma_r}\|M_p - \hat{M}\|_F^2$. *Next, we proceed with the proof by contradiction. First, assume that* $G > \sigma_r(1 + \delta + \zeta_2 q)$. *Then, there exists a small constant* c *such that:*

$$\langle \nabla_M f(\hat{M}, w), M_p - \hat{M}\rangle < -\frac{(1 + \delta + \zeta_2 q) + c}{2}\|M_p - \hat{M}\|_F^2. \tag{7.15}$$

Second, combining the Taylor expansion of $f(M, w)$ *in terms of* M *at the point* $\hat{M}$ *with the mean-value theorem gives:*

$$\begin{aligned}f(M_p, w) =&f(\hat{M}, w) + \langle \nabla_M f(\hat{M}, w), M_p - \hat{M}\rangle + \\ &\frac{1}{2}[\nabla^2 f(\tilde{M}, w)](M_p - \hat{M}, M_p - \hat{M}),\end{aligned}$$

for some matrix $\tilde{M}$ *that is a convex combination of* M_p *and* $\hat{M}$. *Due to the RIP assumption and* (7.4), *we have:*

$$\begin{aligned}f(M_p, w) \leq&f(\hat{M}, w) + \langle \nabla_M f(\hat{M}, w), M_p - \hat{M}\rangle + \\ &\frac{1}{2}[(1 + \delta + \zeta_2 q) + c]\|M_p - \hat{M}\|_F^2,\end{aligned} \tag{7.16}$$

for the same small constant c used above. Therefore, by combining (7.15) *and* (7.16)*, we have:*

$$f(M_p, w) < f(\hat{M}, w),$$

which is a contradiction due to the fact that $\hat{X}$ *is a local minimum since we can adjust* p *to make* M_p *arbitrarily close to* $\hat{M} = \hat{X}\hat{X}^\top$ *and that* M_p *is a positive semidefinite matrix of rank* r*. This further leads to the conclusion that* $G \leq \sigma_r(1 + \delta + \zeta_2 q)$*, consequently leading to* (7.14)*.*

Then consider the case where $\operatorname{rank}(\hat{M}) < r$*. By [28], we know that* $\hat{M}$ *is a critical point of* (7.1)*, meaning that if* $\operatorname{rank}(\hat{M}) < r$*,* $\nabla_M f(\hat{M}, w) = 0$*. Therefore* $G = 0$ *and* (7.14) *is trivially satisfied since* $\lambda_r(\hat{M}) = 0$*.*

Proof 35 (Proof of Theorem 16) *Define* $\hat{M} := \hat{X}\hat{X}^\top$ *and*

$$\bar{M} := \hat{M} - \frac{1}{1 + \delta + \zeta_2 q} \nabla_M f(\hat{M}, w). \tag{7.17}$$

Additionally, define $\phi(\cdot)$ *as*

$$\phi(M) := \langle \nabla_M f(\hat{M}, w), M - \hat{M} \rangle + \frac{1 + \delta + \zeta_2 q}{2} \|M - \hat{M}\|_F^2.$$

Now,

$$\frac{1 + \delta + \zeta_2 q}{2} \|M - \bar{M}\|_F^2 = \frac{1 + \delta + \zeta_2 q}{2} \|M - \hat{M} + \frac{1}{1 + \delta + \zeta_2 q} \nabla_M f(\hat{M}, w)\|_F^2$$

$$= \frac{1 + \delta + \zeta_2 q}{2} \|M - \hat{M}\|_F^2 + \langle \nabla_M f(\hat{M}, w), M - \hat{M} \rangle + \frac{1}{(1 + \delta + \zeta_2 q)^2} \|\nabla_M f(\hat{M}, w)\|_F^2$$

$$= \phi(M) + \textit{constant with respect to } M.$$

Define $\mathcal{P}_r(M)$ *of an arbitrary matrix* M *to be the projection of* M *on a low-rank manifold of rank at most* r*:*

$$\mathcal{P}_r(M) = \operatorname*{arg\,min}_{M_r \in \mathcal{M}} \|M_r - M\|_F, \qquad \mathcal{M} := \{M \in \mathbb{S}^{n \times n} | \operatorname{rank}(M) \leq r, M \succeq 0\}$$

Then by the Eckart-Young-Mirsky Theorem, $\phi(\mathcal{P}_r(\bar{M}))$ *achieves the minimum value of the function* $\phi(\cdot)$ *over all matrices of rank at most* r*. Therefore,*

$$-\phi(\mathcal{P}_r(\bar{M})) \geq -\phi(M^*) = \langle \nabla_M f(\hat{M}, w), \hat{M} - M^* \rangle - \frac{1 + \delta + \zeta_2 q}{2} \|M^* - \hat{M}\|_F^2. \tag{7.18}$$

Next, we apply the Taylor expansion to $f(M, w)$ *at* $\hat{M}$ *and combine it with the RIP property to obtain*

$$f(M^*, w) \geq f(\hat{M}, w) + \langle \nabla_M f(\hat{M}, w), M^* - \hat{M} \rangle + \frac{1 - \delta - \zeta_2 q}{2} \|M^* - \hat{M}\|_F^2. \tag{7.19}$$

Additionally, by expanding at M^**, we can also write:*

$$\begin{aligned} f(\hat{M}, w) - f(M^*, w) &\geq \langle \nabla_M f(M^*, w), \hat{M} - M^* \rangle + \frac{1 - \delta - \zeta_2 q}{2} \|\hat{M} - M^*\|_F^2 \\ &\geq \frac{1 - \delta - \zeta_2 q}{2} \|\hat{M} - M^*\|_F^2 - \zeta_1 q \|\hat{M} - M^*\|_F \end{aligned} \tag{7.20}$$

where the second inequality follows from (7.3) *and the fact that* M^* *is the ground truth. Substituting* (7.19) *into* (7.18) *gives:*

$$-\phi(\mathcal{P}_r(\bar{M})) \geq f(\hat{M}, w) - f(M^*, w) - (\delta + \zeta_2 q) \|\hat{M} - M^*\|_F^2,$$

and a further substitution of (7.20) *into the above equation gives:*

$$-\phi(\mathcal{P}_r(\bar{M})) \geq \frac{1 - 3\delta - 3\zeta_2 q}{2} \|\hat{M} - M^*\|_F^2 - \zeta_1 q \|\hat{M} - M^*\|_F. \tag{7.21}$$

We denote

$$L := \frac{1 - 3\delta - 3\zeta_2 q}{2} \|\hat{M} - M^*\|_F^2 - \zeta_1 q \|\hat{M} - M^*\|_F.$$

Next, for the notational simplicity of the ensuing sections, define:

$$N := -\frac{1}{1 + \delta + \zeta_2 q} \nabla f(\hat{M}, w),$$

implying that $\bar{M} = \hat{M} + N$*. Then,*

$$\begin{aligned} -\phi(\mathcal{P}_r(\bar{M})) &= (1 + \delta + \zeta_2 q) \langle N, \mathcal{P}_r(\hat{M} + N) - \hat{M} \rangle - \frac{1 + \delta + \zeta_2 q}{2} \|\mathcal{P}_r(\hat{M} + N) - \hat{M}\|_F^2 \\ &= \frac{1 + \delta + \zeta_2 q}{2} (\|N\|_F^2 - \|\hat{M} + N - \mathcal{P}_r(\hat{M} + N)\|_F^2) \\ &= \frac{1 + \delta + \zeta_2 q}{2} (\|N\|_F^2 - \|\hat{M} + N\|_F^2 + \|\mathcal{P}_r(\hat{M} + N)\|_F^2). \end{aligned}$$

Since $\hat{X}$ *is a local minimizer of* (7.2)*, it must be a first-order critical point. Therefore,* (2.5) *holds true, meaning that* $\hat{M}$ *and* N *have orthogonal column/row spaces, leading to* $\|\hat{M} + N\|_F^2 = \|\hat{M}\|_F^2 + \|N\|_F^2$*.*

Furthermore, due to the orthogonal nature of $\hat{M}$ *and* N*,* $\|\mathcal{P}_r(\hat{M}+N)\|_F^2$ *is simply the sum of the squares of the maximal* r *eigenvalues of* $\hat{M}$ *and* N *combined, which we assume to be* $\lambda_i(\hat{M}), i \in \{1, \dots, k\}$ *and* $\lambda_i(N), i \in \{1, \dots, r - k\}$*. Therefore,*

$$\|\mathcal{P}_r(\hat{M} + N)\|_F^2 = \sum_{i=1}^{k} \lambda_i(\hat{M})^2 + \sum_{i=1}^{r-k} \lambda_i(N)^2.$$

Subsequently,

$$
\begin{aligned}
-\phi(\mathcal{P}_r(\bar{M})) &= \frac{1+\delta+\zeta_2 q}{2}(-\sum_{i=1}^{r}\lambda_i(\hat{M})^2+\sum_{i=1}^{k}\lambda_i(\hat{M})^2+\sum_{i=1}^{r-k}\lambda_i(N)^2) \\
&= \frac{1+\delta+\zeta_2 q}{2}(-\sum_{i=k+1}^{r}\lambda_i(\hat{M})^2+\sum_{i=1}^{r-k}\lambda_i(N)^2) \\
&\leq \frac{1+\delta+\zeta_2 q}{2}(-(r-k)\lambda_r^2(\hat{M})+(r-k)\lambda_{\max}^2(N)).
\end{aligned}
$$

Then invoking (7.21) *gives:*

$$
(r-k)\lambda_r^2(\hat{M}) \leq -\frac{2L}{1+\delta+\zeta_2 q}+(r-k)\frac{G^2}{(1+\delta+\zeta_2 q)^2}, \tag{7.22}
$$

where $G=-\lambda_{\min}(\nabla f(\hat{M},w))$.

First, assume that $k<r$ *and* $\lambda_r(\hat{M})>0$. *We have*

$$
\lambda_r^2(\hat{M}) \leq \frac{G^2}{(1+\delta+\zeta_2 q)^2}-\frac{2}{1+\delta+\zeta_2 q}\frac{L}{r-k}, \tag{7.23}
$$

Now, recall Lemma 28, which also holds for all local minimizers $\hat{X}$. *A necessary and sufficient condition for both Lemma 28 and* (7.23) *to hold is that:*

$$
L \leq 0, \tag{7.24}
$$

subsequently meaning that,

$$
(1-3\delta-3\zeta_2 q)\|\hat{M}-M^*\|_F^2-2\zeta_1 q\|\hat{M}-M^*\|_F \leq 0
$$

which directly gives (7.5) *after simple rearrangements.*

In the case that $k=r$ *or* $\lambda_r(\hat{M})=0$, (7.22) *reduces to* (7.24) *as well, leading to the same result presented in* (7.5).

7.A.2 Proof of Section 7.2.2

Given a matrix $\hat{X}$, we aim to find the smallest δ such that there is an instance of the problem with this RIP constant for which $\hat{X}$ is a local minimizer that is not associated with the ground truth. For notational convenience, we denote this optimal value as $\delta^*(\hat{X})$. Namely, $\delta^*(\hat{X})$ is the optimal value to the following optimization problem:

$$
\begin{aligned}
\min_{\delta, f(\cdot,w)} \quad & \delta \\
\text{s.t.} \quad & \hat{X} \text{ is a local minimizer of } f(\cdot,w), \\
& f(\cdot,0) \text{ satisfies the } \delta\text{-RIP}_{2r} \text{ property.}
\end{aligned} \tag{7.25}
$$

By the above optimization problem, we know that $\delta \geq \delta^*(\hat{X})$ for all local minimizers $\hat{X}$ of $f(\cdot, w)$, where δ is the best RIP constant of the problem. Since (7.25) is difficult to analyze, we replace its two constraints with some necessary conditions, thus forming a relaxation of the original problem with its optimal value being a lower bound on $\delta^*(\hat{X})$.

To find a necessary condition replacing the two constraints, we introduce the following lemma. This is the first lemma that captures the necessary conditions of a critical point of (7.2), a problem where random noise is considered.

Lemma 29 *Assume that the objective function $f(M, w)$ of* (7.2) *satisfies all assumptions in Section 7.1.2, and that $\hat{X}$ is a first-order critical point of* (7.2). *Then, $\hat{X}$ must satisfy the following conditions for some symmetric matrix $\mathbf{H} \in \mathbb{R}^{n^2 \times n^2}$:*

1. $\|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq 2\zeta_1 q \|\hat{X}\|_2$

2. $\mathbf{H}$ *satisfies the $(\delta + \zeta_2 q)$-RIP$_{2r,2r}$ property, which means that the inequality*

$$(1 - \delta - \zeta_2 q)\|M\|_F^2 \leq \mathbf{m}^\top \mathbf{H}\mathbf{m} \leq (1 + \delta + \zeta_2 q)\|M\|_F^2 \tag{7.26}$$

holds for every matrix $M \in \mathbb{R}^{n \times n}$ with $\operatorname{rank}(M) \leq 2r$, *where* $\mathbf{m} = \operatorname{vec}(M)$ *and* $\mathbf{e} = \operatorname{vec}(\hat{X}\hat{X}^\top - M^*)$. $\hat{\mathbf{X}}$ *is defined as per Section 2.1.*

Given Lemma 29, we can obtain a relaxation of problem (7.25), namely the following optimization problem:

$$\begin{aligned} \min_{\delta, \mathbf{H}} \quad & \delta \\ \text{s. t.} \quad & \|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq 2\zeta_1 q \|\hat{X}\|_2, \\ & (1 - \delta - \zeta_2 q)\|M\|_F^2 \leq \mathbf{m}^\top \mathbf{H}\mathbf{m} \leq \\ & (1 + \delta + \zeta_2 q)\|M\|_F^2, \quad \forall M : \operatorname{rank}(M) \leq 2r. \end{aligned} \tag{7.27}$$

where $\mathbf{m} = \operatorname{vec}(M)$. Note that since the second constraint is hard to deal with, so we solve the following problem that has the same optimal value (as proved in Lemma 14 of [6]):

$$\begin{aligned} \min_{\delta, \mathbf{H}} \quad & \delta \\ \text{s. t.} \quad & \|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq 2\zeta_1 q \|\hat{X}\|_2, \\ & (1 - \delta - \zeta_2 q) I_{n^2} \preceq \mathbf{H} \preceq (1 + \delta + \zeta_2 q) I_{n^2}. \end{aligned} \tag{7.28}$$

If the optimal value of (7.28) is denoted as $\delta_f^*(\hat{X})$, then we know that $\delta_f^*(\hat{X}) \leq \delta^*(\hat{X}) \leq \delta$ due to (7.27) being a relaxation of (7.25). By further lower-bounding $\delta_f^*(\hat{X})$ with an expression in terms of $\|\hat{X}\hat{X}^\top - M^*\|_F$, we can obtain an upper bound on $\|\hat{X}\hat{X}^\top - M^*\|_F$.

Proof 36 (Proof of Lemma 29) *Similar to the last section, we first define $\hat{M} = \hat{X}\hat{X}^\top$. Since $\hat{X}$ is a first-order critical point, it follows from* (2.5) *that $\nabla_X h(\hat{X}, w) = 0$. Thus,*

$$0 = \langle \nabla_X h(\hat{X}, w), U \rangle = \langle \nabla_M f(\hat{M}, w), \hat{X}U^\top + U\hat{X}^\top \rangle, \tag{7.29}$$

for an arbitrary $U \in \mathbb{R}^{n\times r}$. Let $u = \operatorname{vec}(U)$.

Next, we define the function $g(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}$:

$$g(V) = \langle \nabla_M f(V, w), \hat{X}U^\top + U\hat{X}^\top \rangle,$$

for all $V \in \mathbb{R}^{n\times n}$. Then, $g(\hat{M}) = 0$ due to (7.29).

By the mean-value theorem (MTV), we have:

$$\begin{aligned} g(\hat{M}) - g(M^*) &= \int_0^1 \langle \nabla g(tM^* + (1-t)\hat{M}), \hat{M} - M^* \rangle \mathrm{d}t \\ &= \int_0^1 [\nabla_M^2 f(tM^* + (1-t)\hat{M})](\hat{M} - M^*, \hat{X}U^\top + U\hat{X}^\top)\mathrm{d}t \\ &= \mathbf{e}^\top \mathbf{H}\hat{\mathbf{X}}u \end{aligned}$$

where $\mathbf{H} \in \mathbb{R}^{n^2\times n^2}$ is a symmetric matrix that is independent of U and satisfies:

$$\operatorname{vec}(K)^\top \mathbf{H} \operatorname{vec}(L) = \int_0^1 [\nabla_M^2 f(tM^* + (1-t)\hat{M})](K, L)\mathrm{d}t$$

for all $K, L \in \mathbb{R}^{n\times n}$. This means:

$$\mathbf{e}^\top \mathbf{H}\hat{\mathbf{X}}u = g(\hat{M}) - g(M^*).$$

Taking the absolute value of both sides and upper-bounding the right-hand side gives:

$$\begin{aligned} |\mathbf{e}^\top \mathbf{H}\hat{\mathbf{X}}u| = |g(\hat{M}) - g(M^*)| &\leq |g(M^*)| \\ &\leq \zeta_1 q \|\hat{X}U^\top + U\hat{X}^\top\|_F \\ &\leq 2\zeta_1 q \|\hat{X}U^\top\|_F \\ &= 2\zeta_1 q \sqrt{\operatorname{tr}(\hat{X}\hat{X}^\top UU^\top)} \\ &\leq 2\zeta_1 q \|\hat{X}\|_2 \|u\|, \end{aligned}$$

where the second line follows from combining (7.13) *and* (7.3), *and the fourth line follows from the cyclic property of trace operators.*

Choosing $u = \hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}$ can simplify the above inequality to

$$\|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq 2\zeta_1 q \|\hat{X}\|_2.$$

Furthermore, the δ-RIP$_{2r,2r}$ property of the objective function means that:

$$(1-\delta)\|M\|_F^2 \le [\nabla^2 f(\xi,0)](M,M) \le (1+\delta)\|M\|_F^2$$

for all M with $\operatorname{rank}(M) \le 2r$. *Combining with the fact that*

$$|\operatorname{vec}(M)^\top \mathbf{H} \operatorname{vec}(M) - [\nabla^2 f(\xi,0)](M,M)| \le \zeta_2 q \|M\|_F^2,$$

gives (7.26).

Proof 37 (Proof of Theorem 17) *One can replace the decision variable δ in* (7.28) *with η and introduce the following optimization problem:*

$$\begin{aligned} \max_{\eta,\hat{\mathbf{H}}} \quad & \eta \\ \text{s.t.} \quad & \|\hat{\mathbf{X}}^\top \hat{\mathbf{H}} \mathbf{e}\| \le 2\zeta_1 q \|\hat{X}\|_2, \\ & \eta I_{n^2} \preceq \hat{\mathbf{H}} \preceq I_{n^2}. \end{aligned} \tag{7.30}$$

It is easy to realize that given any feasible solution $(\delta, \mathbf{H})$ for (7.28)*, the following pair of points will serve as a feasible solution to* (7.30)*:*

$$\eta = \frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q}, \qquad \hat{\mathbf{H}} = \frac{1}{1+\delta+\zeta_2 q}\mathbf{H}.$$

By denoting the optimal value of (7.30) *as $\eta_f^*(\hat{X})$, it holds that*

$$\eta_f^*(\hat{X}) \ge \frac{1-\delta_f^*(\hat{X})-\zeta_2 q}{1+\delta_f^*(\hat{X})+\zeta_2 q} \ge \frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q}, \tag{7.31}$$

for all local minimizers (it is important to recall $\delta_f^*(\hat{X}) \le \delta^*(\hat{X}) \le \delta$).

As stated above, the key to proving (7.7) *is to upper-bounding $\eta_f^*(\hat{X})$. Since* (7.30) *is a semidefinite programming problem, finding any feasible solution of its Lagrangian dual can provide an upper bound. The dual problem is given as follows:*

$$\begin{aligned} \min_{U_1,U_2,G,\lambda,y} \quad & \operatorname{tr}(U_2) + 4\zeta_1^2 q^2 \|\hat{X}\|_2^2 \lambda + \operatorname{tr}(G) \\ \text{s.t.} \quad & \operatorname{tr}(U_1) = 1, \\ & (\hat{\mathbf{X}} y)\mathbf{e}^\top + \mathbf{e}(\hat{\mathbf{X}} y)^\top = U_1 - U_2, \\ & \begin{bmatrix} G & -y \\ -y^\top & \lambda \end{bmatrix} \succeq 0, \\ & U_1 \succeq 0, \quad U_2 \succeq 0. \end{aligned} \tag{7.32}$$

As per Chapter 6, define

$$M = (\hat{\mathbf{X}}y)\mathbf{e}^\top + \mathbf{e}(\hat{\mathbf{X}}y)^\top,$$

and decompose M *as* $M = [M]_+ - [M]_-$ *with* $[M]_+ \succeq 0$ *and* $[M]_- \succeq 0$*. Then, we find a set of feasible solutions* $(U_1^*, U_2^*, G^*, \lambda^*, y^*)$ *to (7.32), which are:*

$$y^* = \frac{y}{\operatorname{tr}([M]_+)}, \quad U_1^* = \frac{[M]_+}{\operatorname{tr}([M]_+)}, \quad U_2^* = \frac{[M]_-}{\operatorname{tr}([M]_+)},$$
$$G^* = \frac{y^*(y^*)^\top}{\lambda^*}, \quad \lambda^* = \frac{\|y^*\|}{2\zeta_1 q\|\hat{X}\|_2}.$$

It is easy to verify that the above solution is feasible and has the objective value

$$\frac{\operatorname{tr}([M]_-) + 4\zeta_1 q\|\hat{X}\|_2\|y\|}{\operatorname{tr}([M]_+)}. \tag{7.33}$$

For any matrix $\hat{X} \in \mathbb{R}^{n\times r}$ *satisfying* $\|\hat{X}\hat{X}^\top - M^*\|_F \le \tau\lambda_r(M^*)$*, we have* $\hat{X} \neq 0$*. Moreover, it has been shown in the proof of Lemma 19 in [6] that any* $y \neq 0$ *for which* $\hat{X}^\top \operatorname{mat}(y)$ *is symmetric satisfies the inequality*

$$\|\hat{\mathbf{X}}y\|^2 \ge 2\lambda_r(\hat{X}\hat{X}^\top)\|y\|^2. \tag{7.34}$$

where r *is the rank of* $\hat{X}$*. Furthermore, by the Wielandt–Hoffman theorem,*

$$|\lambda_r(\hat{X}\hat{X}^\top) - \lambda_r(M^*)| \le \|\hat{X}\hat{X}^\top - M^*\|_F \le \tau\lambda_r(M^*),$$
$$|\lambda_1(\hat{X}\hat{X}^\top) - \lambda_1(M^*)| \le \|\hat{X}\hat{X}^\top - M^*\|_F \le \tau\lambda_r(M^*).$$

Thus, using the above two inequalities and (7.34), we have

$$\frac{2\|\hat{X}\|_2\|y\|}{\|\hat{\mathbf{X}}y\|} \le \frac{2\|\hat{X}\|_2}{\sqrt{2\lambda_r(\hat{X}\hat{X}^\top)}} \le \sqrt{\frac{2(\lambda_1(M^*) + \tau\lambda_r(M^*))}{(1-\tau)\lambda_r(M^*)}} := C(\tau, M^*). \tag{7.35}$$

The second inequality holds because

$$\begin{aligned}\lambda_r(\hat{X}\hat{X}^\top) &= \lambda_r(M^*) - (\lambda_r(M^*) - \lambda_r(\hat{X}\hat{X}^\top)) \\ &\ge \lambda_r(M^*) - |(\lambda_r(M^*) - \lambda_r(\hat{X}\hat{X}^\top)| \\ &\ge \lambda_r(M^*) - \tau\lambda_r(M^*) = (1-\tau)\lambda_r(M^*)\end{aligned}$$

Next, according to Lemma 14 of [101], one can write

$$\operatorname{tr}([M]_+) = \|\hat{\mathbf{X}}y\|\|\mathbf{e}\|(1+\cos\theta),$$
$$\operatorname{tr}([M]_-) = \|\hat{\mathbf{X}}y\|\|\mathbf{e}\|(1-\cos\theta).$$

where θ is the angle between $\hat{\mathbf{X}}y$ and $\mathbf{e}$. Substituting the above two equations and (7.35) *into the dual objective value* (7.33)*, one can obtain*

$$\eta_f^*(\hat{X}) \leq \frac{1 - \cos\theta + 2\zeta_1 q C(\tau, M^*)/\|\mathbf{e}\|}{1 + \cos\theta},$$

which together with (7.31) *implies that*

$$\|\mathbf{e}\| \leq \frac{(1 + \delta + \zeta_2 q)\zeta_1 q C(\tau, M^*)}{\cos\theta - \zeta_2 q - \delta}. \tag{7.36}$$

Now, we seek to lower-bound $\cos(\theta)$*. This amounts to taking the upper bound of* $\sin^2(\theta)$*. This requires us to choose a particular value of y. We choose the same y that is described in Lemma 12 of [93], since it makes $\hat{X}^\top \operatorname{mat}(y)$ symmetric, thereby satisfying* (7.34)*. From the proof of Lemma 13 of [93], we know:*

$$\sin^2(\theta) = \frac{\|Z^\top(I - \hat{X}\hat{X}^\dagger)Z\|_F^2}{\|\hat{X}\hat{X}^\top - ZZ^\top\|_F^2},$$

Since the expression of $\sin^2(\theta)$ *is invariant to re-scaling, we may re-scale both $\hat{X}$ and Z until $\|ZZ^\top\|_F^2 = 1$. Also, since the expression is rotationally invariant, we can partition $\hat{X}$ and Z as follows:*

$$\hat{X} = \begin{bmatrix} X_1 \\ 0 \end{bmatrix} \qquad Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

where $X_1, Z_1 \in \mathbb{R}^{r\times r}, Z_2 \in \mathbb{R}^{(n-r)\times r}$. We compute the QR decomposition $QR = [X, Z]$ and redefine $X := Q^\top X, Z := Q^\top Z$. Then, we follow the technique in Lemma 13 to arrive at:

$$\frac{\|Z^\top(I - \hat{X}\hat{X}^\dagger)Z\|_F^2}{\|\hat{X}\hat{X}^\top - ZZ^\top\|_F^2} = \frac{\|Z_2(Z_2)^\top\|_F^2}{\|Z_1(Z_1)^\top - X_1X_1^\top\|_F^2 + 2\|Z_1(Z_2)^\top\|_F^2 + \|Z_2(Z_2)^\top\|_F^2}.$$

Additionally,

$$\begin{aligned} \sigma_{\min}^2(Z_1) =& \lambda_{\min}((Z_1)^\top(Z_1)) \\ \geq& \lambda_{\min}((Z_1)^\top(Z_1) + (Z_2)^\top(Z_2)) - \lambda_{\max}((Z_2)^\top Z_2) \\ =& \sigma_r^2(Z) - \|Z_2(Z_2)^\top\|_2 \\ \geq& \sigma_r^2(Z) - \tau\lambda_r(M^*) = (1-\tau)\lambda_r(M^*). \end{aligned} \tag{7.37}$$

The last line of (7.37) *is due to*

$$\begin{aligned} \tau^2\lambda_r^2(M^*) \geq& \|\hat{X}\hat{X}^\top - ZZ^\top\|_F^2 \\ =& \|Z_1(Z_1)^\top - X_1X_1^\top\|_F^2 + 2\|Z_1(Z_2)^\top\|_F^2 + \|Z_2(Z_2)^\top\|_F^2 \\ \geq& \|Z_2(Z_2)^\top\|_F^2, \end{aligned} \tag{7.38}$$

and that $\|Z_2(Z_2)^\top\|_F \geq \|Z_2(Z_2)^\top\|_2$.

Subsequently,

$$\begin{aligned}
\sin^2(\theta) &\leq \frac{\|Z_2(Z_2)^\top\|_F^2}{2\|Z_1(Z_2)^\top\|_F^2 + \|Z_2(Z_2)^\top\|_F^2} \\
&\leq \frac{\|Z_2(Z_2)^\top\|_F\|Z_2\|_F^2}{2\sigma_{\min}^2(Z_1)\|Z_2\|_F^2 + \|Z_2(Z_2)^\top\|_F\|Z_2\|_F^2} \\
&\leq \frac{\|Z_2(Z_2)^\top\|_F}{2(1-\tau)\lambda_r(M^*) + \|Z_2(Z_2)^\top\|_F} \\
&\leq \frac{\tau\lambda_r(M^*)}{2(1-\tau)\lambda_r(M^*) + \tau\lambda_r(M^*)} \\
&\leq \frac{\tau}{(2-\tau)} \leq \tau,
\end{aligned}$$

where the first inequality follows from the fact that $\|Z_1(Z_1)^\top - X_1X_1^\top\|_F^2 \geq 0$, *the third inequality follows from* (7.37), *and the fourth inequality follows from* (7.38) *and the fact that the function* $\frac{x}{c+x}$ *is increasing with* x *when both* c *and* x *are positive.*

The above bound is automatically non-vacuous, since $\sin^2(\theta) \leq \tau < 1$. *Therefore,*

$$\cos\theta \geq \sqrt{1-\tau},$$

leading to (7.7) *after substitution into* (7.36).

7.A.3 Proof of Section 7.3.1

First and foremost, we restate this lemma from [78, 105]:

Lemma 30 *For any matrix* $X \in \mathbb{R}^{n\times r}$, *given a positive semidefinite matrix* $M \in \mathbb{R}^{n\times n}$ *of rank* r, *we have:*

$$\|XX^\top - M\|_F^2 \geq 2(\sqrt{2}-1)\sigma_r(M)(\text{dist}(X, M))^2. \tag{7.39}$$

Also, given Assumption 5, we have

$$\nabla_M f(M^w, w) = 0 \tag{7.40}$$

First, we establish that the PL inequality holds in a neighborhood of the global minimizer.

Lemma 31 *Consider the global minimizer* M^w *of* (7.1). *There exists a constant* $\mu > 0$ *such that the PL inequality:*

$$\frac{1}{2}\|\nabla_X h(X, w)\|_F^2 \geq \mu(h(X, w) - f(\mathcal{P}_r(M^w), w)), \tag{7.41}$$

holds for all $X \in \mathbb{R}^{n\times r}$ satisfying:

$$\mathrm{dist}(X, M^w) < \max\{\sqrt{2(\sqrt{2}-1)}\sqrt{1-(\delta+\zeta_2 q)^2}(\sigma_r(M^w))^{1/2} - D_r, 0\} \tag{7.42}$$

and

$$D_r \leq \mathrm{dist}(X, \mathcal{P}_r(M^w)),$$

for $q < (1-\delta)/\zeta_2$.

Proof 38 (Proof of Lemma 31) *We prove the Lemma when $C_w\sqrt{1-(\delta+\zeta_2 q)^2} - D_r > 0$, since otherwise it is trivial. Denote $M := XX^\top$. First, we fix a constant $\tilde{C}$ such that:*

$$\mathrm{dist}(X, M^w) \leq \tilde{C} < C_w\sqrt{1-(\delta+\zeta_2 q)^2} - D_r. \tag{7.43}$$

Then, we define q_1 and q_2 as follows:

$$q_1 = \sqrt{1 - \frac{\tilde{C}^2}{2(\sqrt{2}-1)\sigma_r(M^w)}}, q_2 = \frac{\sqrt{2}\mu'}{\sigma_r(M^w)^{1/2} - \tilde{C}}. \tag{7.44}$$

Now, both q_1 and q_2 are nonnegative resulting from the assumption above. Furthermore, we know that $\delta + \zeta_2 q < \sqrt{1 - \frac{\tilde{C}^2}{2(\sqrt{2}-1)\sigma_r(M^w)}}$ from (7.43), then

$$\frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q} > \frac{1-q_1+q_2}{1+q_1}, \tag{7.45}$$

for some small enough μ'. Define $\mu = (\mu')^2/(1+\delta+\zeta_2 q + 2\rho)$. First, we make the assumption that:

$$\frac{1}{2}\|\nabla_X h(X,w)\|_F^2 < \mu(h(X,w) - f(\mathcal{P}_r(M^w), w)). \tag{7.46}$$

From this assumption, we have:

$$\begin{aligned}
&\mu(h(X,w) - f(\mathcal{P}_r(M^w), w)) \\
\leq&\mu\left(\langle \nabla_M f(\mathcal{P}_r(M^w), w), M - \mathcal{P}_r(M^w)\rangle + \frac{1+\delta+\zeta_2 q}{2}\|M - \mathcal{P}_r(M^w)\|_F^2\right) \\
\leq&\mu\left(\rho\|M^w - \mathcal{P}_r(M^w)\|_F\|M - \mathcal{P}_r(M^w)\|_F + \frac{1+\delta+\zeta_2 q}{2}\|M - \mathcal{P}_r(M^w)\|_F^2\right) \\
\leq&\mu\left(\rho\|M - \mathcal{P}_r(M^w)\|_F^2 + \frac{1+\delta+\zeta_2 q}{2}\|M - \mathcal{P}_r(M^w)\|_F^2\right).
\end{aligned}$$

due to Taylor's theorem and (7.4). So then (7.46) leads to:

$$\frac{1}{2}\|\nabla h(X,w)\|_F^2 < \mu(\frac{(1+\delta+\zeta_2 q)}{2} + \rho)\|M - \mathcal{P}_r(M^w)\|_F^2.$$

Therefore,

$$\|\nabla h(X,w)\|_F \leq \mu' \|M - \mathcal{P}_r(M^w)\|_F.$$

Then consider the following optimization problem:

$$\begin{aligned} \min_{\delta, \mathbf{H}\in\mathbb{S}^{n^2}} \quad & \delta \\ \text{s. t.} \quad & \|\hat{\mathbf{X}}^\top \mathbf{H}\mathbf{e}\| \leq \mu' \|\mathbf{e}\|, \\ & \mathbf{H} \textit{ satisfies the } (\delta+\zeta_2 q)\text{-RIP}_{2r} \textit{ property.} \end{aligned} \tag{7.47}$$

where $\mathbf{e} = \mathrm{vec}(XX^\top - \mathcal{P}_r(M^w))$*. If we denote the optimal value of* (7.47) *as* $\delta^*_f(X,\mu')$*, then* $\delta^*_f(X,\mu') \leq \delta$ *because the constraints of* (7.47) *are necessary conditions for* (7.46)*, according to Lemma 12 of [7]. Therefore,*

$$\frac{1-\delta-\zeta_2 q}{1+\delta+\zeta_2 q} \leq \frac{1-\delta^*_f(X,\mu')-\zeta_2 q}{1+\delta^*_f(X,\mu')+\zeta_2 q}.$$

Moreover, by the same logic of (7.31)*, we know that* $\eta^*_f(X,\mu') \geq \frac{1-\delta^*_f(X,\mu')-\zeta_2 q}{1+\delta^*_f(X,\mu')+\zeta_2 q}$*, where* $\eta^*_f(X,\mu')$ *is the optimal value of the optimization problem:*

$$\begin{aligned} \max_{\eta, \hat{\mathbf{H}}} \quad & \eta \\ \text{s. t.} \quad & \|\hat{\mathbf{X}}^\top \hat{\mathbf{H}}\mathbf{e}\| \leq \mu' \|\mathbf{e}\|, \\ & \eta I_{n^2} \preceq \hat{\mathbf{H}} \preceq I_{n^2}. \end{aligned} \tag{7.48}$$

Lemma 14 of [7] gives:

$$\eta^*_f(X,\mu') \leq \frac{1-q_1+q_2}{1+q_1},$$

therefore making a contradiction to (7.45)*, subsequently proving* (7.41)*.*

Proof 39 (Proof of Theorem 18) *If we certify that:*

$$\frac{\|XX^\top - M^w\|_F}{C_w} < C_w\sqrt{1-(\delta+\zeta_2 q)^2} - D_r \tag{7.49}$$

for any given $X \in \mathbb{R}^{n\times r}$*, then a direct substitution can certify that* (7.42) *holds for* X*, since by Lemma 30,*

$$\mathrm{dist}(X,M) \leq \frac{\|XX^\top - M^w\|_F}{C_w}.$$

Therefore, the certification of (7.49) *means that the PL inequality* (7.41) *holds for this given* X*. Given that* (7.10) *is satisfied, then if this inequality holds:*

$$\|XX^\top - M^w\|_F \leq \sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}\|X_0X_0^\top - M^w\|_F, \tag{7.50}$$

(7.49) *will also hold, because:*

$$\sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}\|X_0X_0^\top - M^w\|_F \leq C_w^2\sqrt{1-(\delta+\zeta_2 q)^2} - C_w D_r.$$

Thus, for the remainder of the proof, we aim to certify that starting from X_0, if we apply the gradient descent algorithm, (7.50) *will be satisfied every step along this trajectory.*

In order to do so, we use Taylor's expansion and (7.40) *to obtain*

$$f(M,w) - f(M^w,w) = \frac{[\nabla^2 f(N,w)](M-M^w, M-M^w)}{2},$$

where N is some convex combination of M and M^w, and $M \in \mathbb{R}^{n\times n}$ is any matrix of rank at most r. In light of the RIP property of the function and (7.4)*, one can write:*

$$\frac{1-\delta-\zeta_2 q}{2}\|M-M^w\|_F^2 \leq f(M,w) - f(M^w,w) \leq \frac{1+\delta+\zeta_2 q}{2}\|M-M^w\|_F^2.$$

This means that if $M_1, M_2 \in \mathbb{R}^{n\times n}$ are two matrices of rank at most r with $f(M_1,w) \leq f(M_2,w)$, then:

$$\|M_1 - M^w\|_F \leq \sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}\|M_2 - M^w\|_F, \qquad (7.51)$$

because $f(M_1,w) - f(M^w,w) \leq f(M_2,w) - f(M^w,w)$.

Thus, one can conclude that $f(X_tX_t^\top, w) \leq f(X_0X_0^\top, w)$ $\forall t$, where X_t denotes the t^{th} step of the gradient descent algorithm starting from X_0. Hence, (7.50) *follows for all X_t.*

Conveniently, Lemma 11 in [7] shows that $f(X_tX_t^\top, 0) \leq f(X_{t-1}X_{t-1}^\top, 0)$ for all $t \geq 0$. However, this result can be extended to:

$$f(X_tX_t^\top, w) \leq f(X_{t-1}X_{t-1}^\top, w),$$

by making

$$1/\eta \geq 12\rho r^{(1/2)}\left(\sqrt{\frac{1+\delta+\zeta_2 q}{1-\delta-\zeta_2 q}}\|X_0X_0^\top - M^w\|_F + \|M^w\|_F\right),$$

since $\nabla f(\cdot, w)$ is now a ρ-Lipschitz continuous function. Given (7.10)*, a sufficient condition to the above inequality is that:*

$$\eta \leq \left(12\rho r^{(1/2)}\left(2(\sqrt{2}-1)\sqrt{(1-(\delta+\zeta_2 q)^2} + \|M^w\|_F\right)\right)^{-1}$$

This finally means that the PL inequality (7.41) *is established for the entire trajectory starting from X_0. Now, applying Theorem 1 in [38] gives:*

$$h(X_t,w) - f(\mathcal{P}_r(M^w),w) \leq (1-\mu\eta)^t(h(X_0,w) - f(\mathcal{P}_r(M^w),w)),$$

which implies a linear convergence as desired.

7.A.4 Proof Sketches in Section 7.3.2

The proof of Theorem 19 is highly similar to that of Theorem 7 in [94], albeit with a number of differences. In this section, we will only highlight the differences, since everything else follows in the same manner.

First and foremost, we replace δ with $\delta + \zeta_2 q$ in all of the proofs since in our noisy formulation, the problem is $(\delta + \zeta_2 q)$-RIP$_{2r,2r}$ instead.

Then, we introduce the following Lemma in lieu of Lemma 6 in [94] since $\nabla_M f(M^*, w) \neq 0$ in the noisy formulation:

Lemma 32 *Given a constant $\epsilon > 0$, an arbitrary $X \in \mathbb{R}^{n\times r}$, and the ground truth solution $M^* \in \mathbb{R}^{n\times n}$ of* (7.1)*, if*

$$\|XX^\top\|_F^2 \geq \max\left\{\frac{2(1+\delta+\zeta_2 q)}{1-\delta-(\zeta_2+\zeta_D)q}\|M^*\|_F^2, (\frac{2\lambda\sqrt{r}}{1-\delta-(\zeta_2+\zeta_D)q})^{4/3}\right\}, \qquad (7.52)$$

then

$$\|\nabla_X h(X,w)\|_F \geq \lambda,$$

where $\zeta_D = \zeta_1/D$ and D is a constant such that

$$D^2 \leq (\frac{2\lambda\sqrt{r}}{1-\delta-(\zeta_2+\zeta_D)q})^{4/3}. \qquad (7.53)$$

Note that such D exists since we first require that $1-\delta-(\zeta_2+\zeta_D)q \geq 0$, meaning that $\frac{q\zeta_1}{1-\delta-q\zeta_2} \leq D$. Moreover, a sufficient condition to (7.53) is that $D \leq (2\lambda\sqrt{r})^{2/3}$, which can be simultaneously satisfied when λ is chosen properly. The introduction of the lower bound D will not affect the remainder of the proof of Theorem 19, since in the later steps, we only require the existence of a constant C such that $\|XX^\top\|_F \leq C^2$ when $\|\nabla_X h(X,w)\|_F \leq \lambda$. Therefore, Lemma 32 perfectly fits this role.

Proof 40 (Proof of Lemma 32) *Denote $M \coloneqq XX^\top$. Using the RIP property and* (7.3)*, we have:*

$$\begin{aligned}\langle \nabla_M f(M), M\rangle &= \int_0^1 [\nabla^2 f(M^* + s(M-M^*), w)][M-M^*, M]\mathrm{d}s + \langle \nabla_M f(M^*, w), M\rangle \\ &\geq (1-\delta-\zeta_2 q)\|M\|_F^2 - (1+\delta+\zeta_2 q)\|M^*\|_F\|M\|_F - \zeta_1 q\|M\|_F \\ &= (1-\delta-\zeta_2 q)\|M\|_F^2 - (1+\delta+\zeta_2 q)\|M^*\|_F\|M\|_F - \zeta_D qD\|M\|_F \\ &\geq (1-\delta-(\zeta_2+\zeta_D)q)\|M\|_F^2 - (1+\delta+\zeta_2 q)\|M^*\|_F\|M\|_F \\ &\geq \frac{1-\delta-(\zeta_2+\zeta_D)q}{2}\|M\|_F^2,\end{aligned}$$

where the second last inequality results from (7.53)*, which implies that $D \leq \|M\|_F$; and the last inequality follows from* (7.52)*. Then combining the fact that $\|X\|_F \leq$*

$\sqrt{r}\|M\|_F^{1/2}$, and $\|\nabla_X h(X,w)\|_F \geq \frac{\langle \nabla h(X,w), X\rangle}{\|X\|_F}$ yields the desired fact that

$$\begin{aligned}\|\nabla_X h(X,w)\|_F \geq \frac{\langle \nabla h(X,w), X\rangle}{\|X\|_F} &= \frac{\langle \nabla_M f(M), M\rangle}{\|X\|_F} \\ &\geq \frac{(1-\delta-(\zeta_2+\zeta_D)q)\|M\|_F^2}{2\sqrt{r}\|M\|_F^{1/2}} \\ &= \frac{1-\delta-(\zeta_2+\zeta_D)q}{2\sqrt{r}}\|M\|_F^{3/2} \\ &\geq \lambda.\end{aligned} \tag{7.54}$$

Then, utilizing Lemma 32, we can prove Lemma 7 in [94] in the same fashion to obtain

$$\begin{aligned}&\langle \nabla_M f(M,w), M^*-M\rangle \\ \leq& -(1-\delta-\zeta_2 q)\|M-M^*\|_F^2 - \langle \nabla_M f(M^*,w), M-M^*\rangle \\ \leq& -(1-\delta-\zeta_2 q)\|M-M^*\|_F^2 + \zeta_1 q\|M-M^*\|_F \\ \leq& -(1-\delta-\zeta_2 q)\|M-M^*\|_F^2 + \zeta_\alpha q(\sqrt{2(\sqrt{2}-1)}(\sigma_r(M^*))^{1/2}\alpha)\|M-M^*\|_F \\ \leq& -(1-\delta-(\zeta_2-\zeta_\alpha)q)\|M-M^*\|_F^2\end{aligned}$$

for any $M \in \mathbb{R}^{n\times n}$ that satisfies the requirements in Lemma 7 of [94]. This is because $\|M-M^*\|_F \geq (\sqrt{2(\sqrt{2}-1)}(\sigma_r(M^*))^{1/2}\alpha)$ by the assumption of α and Lemma 30.

The above change will only affect the constant c in Lemma 7, and the new c will become

$$c = (\sqrt{r}\|M^*\|_F)^{-1}(\sqrt{2}-1)(1-\delta-(\zeta_2-\zeta_\alpha)q)\sigma_r(M^*).$$

Since the exact value of c is irrelevant and we only need to prove its existence, the rest of the proof follows from the existing procedure. Note that $c > 0$ is guaranteed by the assumption of noise in Theorem (19). Therefore, Lemma 7 still holds in the noisy case.

Then, we proceed to show that Lemma 8 in [94] can also be proved similarly, except for one key difference, which is

$$K := (1-3\delta-(3\zeta_2+2\zeta_\alpha)q)(\sqrt{2}-1)\sigma_r(M^*)\alpha^2.$$

To verify this statement, we leverage the inequality

$$-\phi(\bar{M}) \geq f(M,w) - f(M^*,w) - (\delta+\zeta_2 q)\|M-M^*\|_F^2,$$

and furthermore we now have that

$$\begin{aligned} f(M,w)-f(M^*,w) &\geq \langle \nabla_M f(M^*,w), M-M^* \rangle + \frac{1-\delta-\zeta_2 q}{2}\|M-M^*\|_F^2 \\ &\geq \frac{1-\delta-\zeta_2 q}{2}\|M-M^*\|_F^2 - \zeta_1 q\|M-M^*\|_F^2 \\ &\geq \frac{1-\delta-\zeta_2 q}{2}\|M-M^*\|_F^2 - \zeta_\alpha q(\sqrt{2(\sqrt{2}-1)}(\sigma_r(M^*))^{1/2}\alpha)\|M-M^*\|_F^2 \\ &\geq \frac{1-\delta-(\zeta_2+2\zeta_\alpha)q}{2}\|M-M^*\|_F^2 \end{aligned}$$

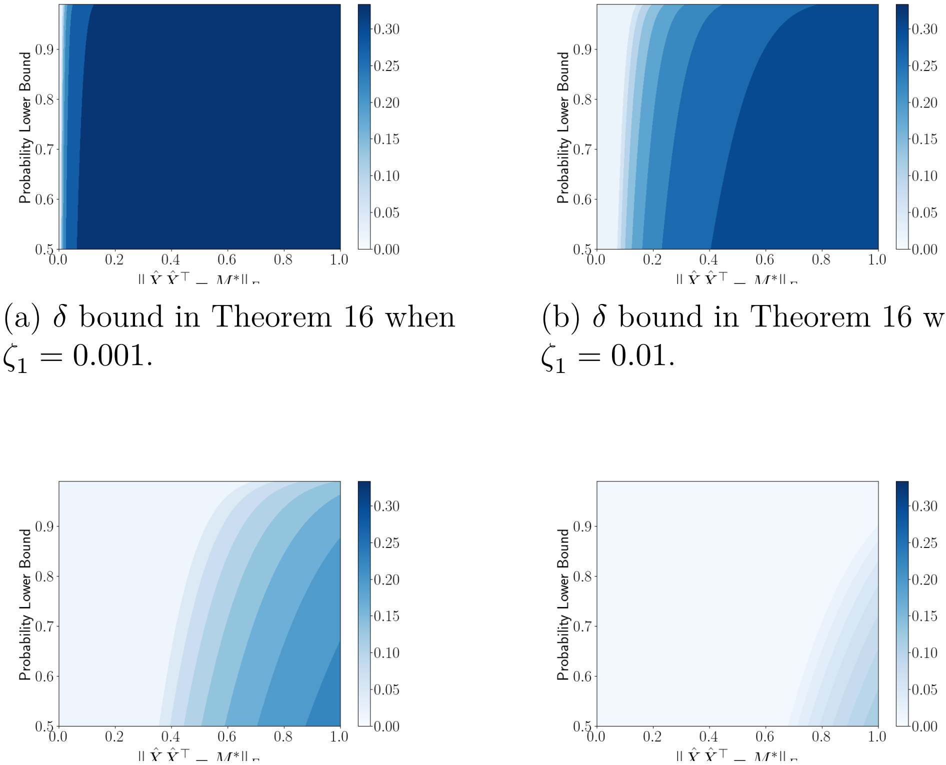
for the same reason elaborated above. Combining the above two inequalities leads to

$$-\phi(\bar{M}) \geq \frac{1-3\delta-(3\zeta_2+2\zeta_\alpha)q}{2}\|M-M^*\|_F^2 \geq K.$$

As assumed in Theorem 19, since $q < \frac{1/3-\delta}{\zeta_2+2\zeta_\alpha/3}$, we know that $K > 0$. This is the only required property of K to facilitate the remainder of the proof of Lemma 8 of [94]. Therefore, Lemma 8 still holds for the noisy case.

Finally, we choose $C = (\frac{2(1+\delta+\zeta_2\epsilon)}{1-\delta-(\zeta_2+\zeta_D)\epsilon}\|M^*\|_F^2)^{1/4}$ and invoke Lemmas 6-8 in [94] to complete the proof of Theorem 19. Note the ϵ here is the same ϵ appeared in the statement of Theorem 19.

7.B Additional Numerical Illustration



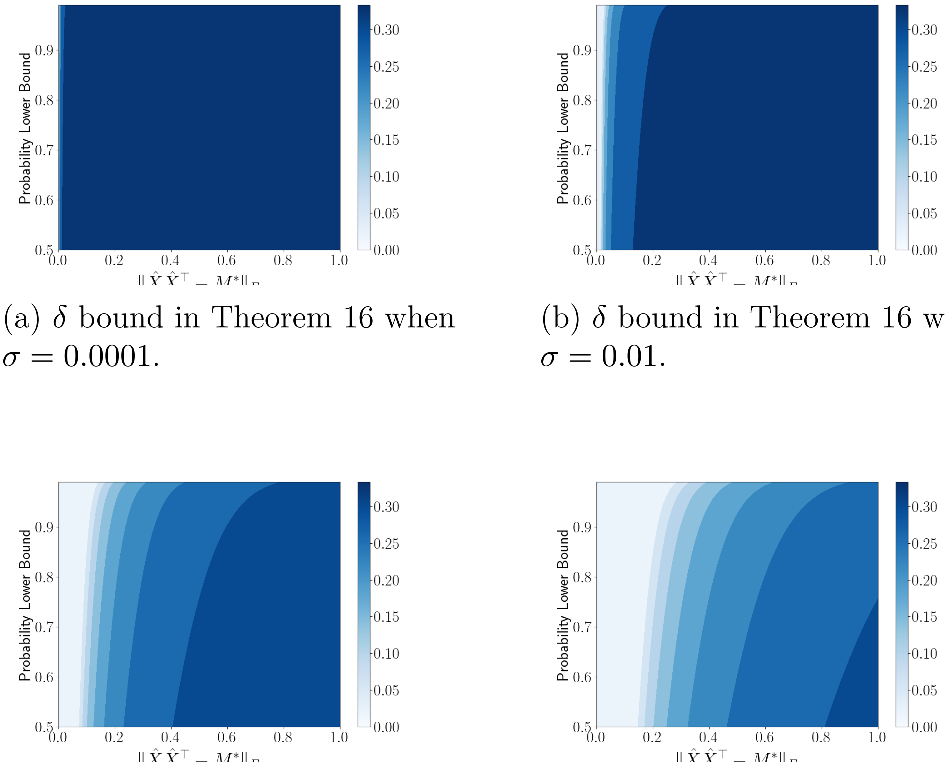
(a) δ bound in Theorem 16 when $\zeta_1 = 0.001$.

(b) δ bound in Theorem 16 when $\zeta_1 = 0.01$.

(c) δ bound in Theorem 16 when $\zeta_1 = 0.05$.

(d) δ bound in Theorem 16 when $\zeta_1 = 0.1$.

Figure 7.3: Comparison of the maximum RIP constants δ allowed by Theorem 16 to guarantee a given bound on the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ for an arbitrary local minimizer $\hat{X}$ satisfying (7.6) with a given probability. In this plot $\zeta_1 = 0, \sigma = 0.05$ as per the main text.



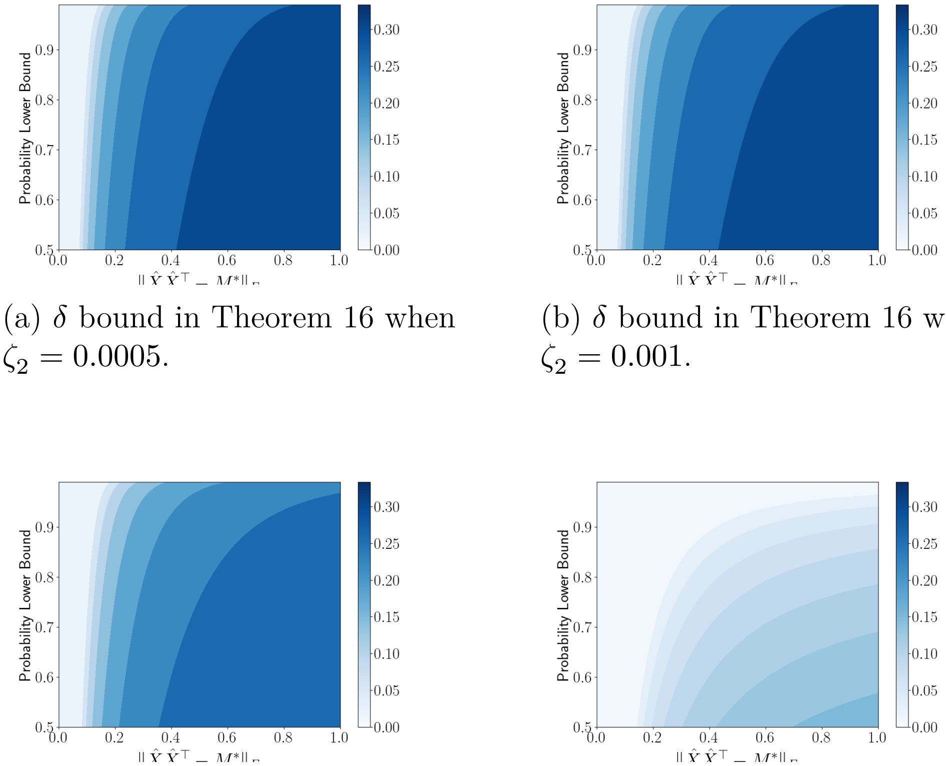
(a) δ bound in Theorem 16 when $\sigma = 0.0001$.

(b) δ bound in Theorem 16 when $\sigma = 0.01$.

(c) δ bound in Theorem 16 when $\sigma = 0.05$.

(d) δ bound in Theorem 16 when $\sigma = 0.1$.

Figure 7.4: Comparison of the maximum RIP constants δ allowed by Theorem 16 to guarantee a given bound on the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ for an arbitrary local minimizer $\hat{X}$ satisfying (7.6) with a given probability. In this plot $\zeta_1 = 0.01, \zeta_2 = 0$.



(a) δ bound in Theorem 16 when $\zeta_2 = 0.0005$.

(b) δ bound in Theorem 16 when $\zeta_2 = 0.001$.

(c) δ bound in Theorem 16 when $\zeta_2 = 0.01$.

(d) δ bound in Theorem 16 when $\zeta_2 = 0.05$.

Figure 7.5: Comparison of the maximum RIP constants δ allowed by Theorem 16 to guarantee a given bound on the distance $\|\hat{X}\hat{X}^\top - M^*\|_F$ for an arbitrary local minimizer $\hat{X}$ satisfying (7.6) with a given probability. In this plot $\zeta_1 = 0.01, \sigma = 0.05$.

Chapter 8

Conclusion

In this dissertation, we embarked on a comprehensive journey to address the non-convex problem of matrix sensing, particularly under two more realistic scenarios: the under-sampled regime and the presence of random noise. Our exploration was rooted in a deep understanding of the current limitations and potential of existing methodologies, leading to the development of novel strategies and theoretical insights that push the boundaries of what is achievable in matrix sensing. The concluding remarks encapsulate our contributions and findings, distilled from the extensive analysis and research conducted throughout this work.

Advancements in the Under-Sampled Regime

Our investigation into the under-sampled regime, characterized by high Restricted Isometry Property (RIP) constants, revealed significant challenges and opportunities for matrix sensing. We highlighted the limitations of traditional approaches like the Semidefinite Programming (SDP) method and the Burer-Monteiro (BM) approach in this context. Through rigorous analysis and novel proofs, we demonstrated improved guarantees for the SDP method, particularly highlighting its enhanced performance with an increasing rank of the solution matrix. This advancement, predicated on the non-existence of incorrect solutions, indicates that SDP can offer robust solutions beyond the previously established bounds, particularly as the RIP constant approaches unity in certain configurations.

Furthermore, our exploration into tensor-based over-parametrization as an alternative to matrix-based solutions opened new avenues for solving matrix sensing problems. By lifting the problem into the tensor space, we not only circumvented the limitations posed by rank constraints but also introduced a methodology for transforming spurious solutions into strict saddle points, thereby enhancing the efficacy of local search methods. This innovative approach underscores the potential of tensor parametrization in tackling non-convex problems more effectively.

Additionally, the introduction of higher-order loss functions emerged as a pivotal development. By incorporating these sophisticated loss functions, we demonstrated the feasibility of escaping from distant critical points, thus offering a viable alternative to over-parametrization. This contribution is particularly relevant in scenarios where practical constraints limit the application of over-parametrization, offering a fresh perspective on navigating the complex landscape of non-convex optimization problems.

Tackling the Noisy Regime

The second focal point of our research addressed the implications of random noise on matrix sensing, a scenario emblematic of real-world data collection and processing challenges. Our work in this domain elucidated the intricate relationship between noise intensity, the RIP constant, and their collective impact on the solution landscape. Through comprehensive theoretical analysis, we established global and local guarantees that delineate the proximity of local minimizers to the ground truth, thereby providing a clearer understanding of the solution's fidelity in the presence of noise.

Moreover, our foray into general low-rank optimization problems in the context of noise further expanded the scope of our findings. By relaxing the RIP constant requirements, we showcased the resilience of our proposed methodologies against noise, reinforcing the notion that accurate recovery is achievable even in less-than-ideal conditions. This part of our research not only broadens the applicability of our findings but also contributes to the broader discourse on low-rank recovery and optimization.

Concluding Remarks

In summary, this dissertation contributes significantly to the field of matrix sensing by providing deeper insights into the challenges and solutions applicable in under-sampled and noisy regimes. Our work extends the theoretical foundations of matrix sensing, offering new perspectives and methodologies that enhance the robustness and applicability of recovery algorithms. By addressing both the under-sampled and noisy scenarios, this research not only advances our understanding of matrix sensing but also lays the groundwork for future explorations in this vibrant field of study. Through this journey, we have not only sought to demystify the complexities inherent in non-convex optimization but also to illuminate pathways towards more effective and efficient problem-solving strategies in the realm of matrix sensing and beyond.

Bibliography

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

[2] 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.

[3] Brendon G Anderson and Somayeh Sojoudi. Global optimality guarantees for nonconvex unsupervised video segmentation. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 965–972. IEEE, 2019.

[4] 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.

[5] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[6] Yingjie Bi and Javad Lavaei. Global and local analyses of nonlinear low-rank matrix recovery problems, 2020. arXiv:2010.04349.

[7] Yingjie Bi, Haixiang Zhang, and Javad Lavaei. Local and global linear convergence of general low-rank matrix recovery problems. *AAAI-22*, 2022.

[8] Nicolas Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377, 2016.

[9] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[10] 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.

[11] T Tony Cai and Anru Zhang. Sharp rip bound for sparse signal and low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 35(1):74–93, 2013.

[12] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[13] Emmanuel J Candes and Yaniv Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *arXiv preprint arXiv:1001.0339*, 2010.

[14] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. 2011.

[15] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[16] 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.

[17] Shih Yu Chang. Hanson-wright inequality for random tensors under einstein product. *arXiv preprint arXiv:2111.12169*, 2021.

[18] 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.

[19] Pierre Comon, Gene Golub, Lek-Heng Lim, and Bernard Mourrain. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279, 2008.

[20] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[21] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[22] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

[23] Salar Fattahi and Somayeh Sojoudi. Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis. *Journal of Machine Learning Research*, 21:1–51, 2020.

[24] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1233–1242, 2017.

[25] Rong Ge, Yunwei Ren, Xiang Wang, and Mo Zhou. Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems*, 34:1299–1311, 2021.

[26] Navid Ghadermarzy, Yaniv Plan, and Ozgur Yilmaz. Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing*, 67(1):29–40, 2018.

[27] Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4):2927–2955, 2020.

[28] Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4):2927–2955, 2020.

[29] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[30] 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.

[31] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1724–1732, August 2017.

[32] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subGaussian norm, 2019. arXiv:1902.03736.

[33] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

[34] Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon S Du, and Jason D Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. *arXiv preprint arXiv:2301.11500*, 2023.

[35] Ming Jin, Javad Lavaei, Somayeh Sojoudi, and Ross Baldick. Boundary defense against cyber threat for power system state estimation. *IEEE Transactions on Information Forensics and Security*, 16:1752–1767, 2021.

[36] Ming Jin, Igor Molybog, Reza Mohammadi-Ghazi, and Javad Lavaei. Towards robust and scalable power system state estimation. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3245–3252. IEEE, 2019.

[37] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[38] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[40] Eleftherios Kofidis and Phillip A Regalia. On the best rank-1 approximation of higher-order supersymmetric tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(3):863–884, 2002.

[41] Tamara G Kolda. Numerical optimization for symmetric tensor decomposition. *Mathematical Programming*, 151(1):225–248, 2015.

[42] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[43] Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

[44] Eitan Levin, Joe Kileel, and Nicolas Boumal. The effect of smooth parametrizations on nonconvex optimization landscapes. *arXiv preprint arXiv:2207.03512*, 2022.

[45] Peiyan Li, Xu Liang, and Haochen Song. A survey on implicit bias of gradient descent. In *2022 14th International Conference on Computer Research and Development (ICCRD)*, pages 108–114. IEEE, 2022.

[46] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019.

[47] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020.

[48] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[49] Lek-Heng Lim and Pierre Comon. Blind multilinear identification. *IEEE Transactions on Information Theory*, 60(2):1260–1280, 2013.

[50] Zhouchen Lin. A review on low-rank models in data analysis. *Big Data and Information Analytics*, 1(2/3):139–161, 2016.

[51] Jianhao Ma and Salar Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *arXiv preprint arXiv:2202.08788*, 2022.

[52] Ziye Ma, Yingjie Bi, Javad Lavaei, and Somayeh Sojoudi. Sharp restricted isometry property bounds for low-rank matrix recovery problems with corrupted measurements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7672–7681, 2022.

[53] Ziye Ma, Yingjie Bi, Javad Lavaei, and Somayeh Sojoudi. Geometric analysis of noisy low-rank matrix recovery in the exact parametrized and the over-parametrized regimes. *INFORMS Journal on Optimization*, 2023.

[54] Ziye Ma, Ying Chen, Javad Lavaei, and Somayeh Sojoudi. Absence of spurious solutions far from ground truth: A low-rank analysis with high-order losses. In *International Conference on Artificial Intelligence and Statistics*, volume 238. PMLR, 2024.

[55] Ziye Ma, Javad Lavaei, and Somayeh Sojoudi. Algorithmic regularization in tensor optimization: Towards a lifted approach in matrix sensing. *Advances in Neural Information Processing Systems*, 36, 2024.

[56] Ziye Ma, Igor Molybog, Javad Lavaei, and Somayeh Sojoudi. Over-parametrization via lifting for low-rank matrix sensing: Conversion of spurious solutions to strict saddle points. In *International Conference on Machine Learning*. PMLR, 2023.

[57] Ziye Ma and Somayeh Sojoudi. Noisy low-rank matrix optimization: Geometry of local minima and convergence rate. In *International Conference on Artificial Intelligence and Statistics*, pages 3125–3150. PMLR, 2023.

[58] Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

[59] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[60] Karthik Mohan and Maryam Fazel. New restricted isometry results for noisy low-rank recovery. In *2010 IEEE International Symposium on Information Theory*, pages 1573–1577. IEEE, 2010.

[61] Igor Molybog, Ramtin Madani, and Javad Lavaei. Conic optimization for quadratic regression under sparse noise. *The Journal of Machine Learning Research*, 21(1):7994–8029, 2020.

[62] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.

[63] Guyan Ni. Hermitian tensor and quantum mixed state. *arXiv preprint arXiv:1902.02640*, 2019.

[64] Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

[65] Dohuyng Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer–Monteiro approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 65–74, 2017.

[66] Pablo A Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical programming*, 96(2):293–320, 2003.

[67] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[68] Liqun Qi. The spectral theory of tensors (rough version). *arXiv preprint arXiv:1201.3424*, 2012.

[69] Liqun Qi, Shenglong Hu, Xinzhen Zhang, and Yannan Chen. Tensor norm, cubic power and gelfand limit. *arXiv preprint arXiv:1909.10942*, 2019.

[70] Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. In *International Conference on Machine Learning*, pages 8913–8924. PMLR, 2021.

[71] Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. In *International Conference on Machine Learning*, pages 18422–18462. PMLR, 2022.

[72] 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.

[73] 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.

[74] Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015.

[75] Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011.

[76] 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*, 34:23831–23843, 2021.

[77] Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2013.

[78] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

[79] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[80] René Vidal, Yi Ma, S Shankar Sastry, René Vidal, Yi Ma, and S Shankar Sastry. Principal component analysis. *Generalized principal component analysis*, pages 25–62, 2016.

[81] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005.

[82] Lingxiao Wang, Xiao Zhang, and Quanquan Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 981–990, 2017.

[83] Alexander S Wein, Ahmed El Alaoui, and Cristopher Moore. The kikuchi hierarchy and tensor pca. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1446–1468. IEEE, 2019.

[84] Fei Wen, Lei Chu, Peilin Liu, and Robert C Qiu. A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access*, 6:69883–69906, 2018.

[85] John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems*, 22, 2009.

[86] Stephen J Wright and Benjamin Recht. *Optimization for data analysis*. Cambridge University Press, 2022.

[87] Leqin Wu, Xin Liu, and Zaiwen Wen. Symmetric rank-1 approximation of symmetric high-order tensors. *Optimization Methods and Software*, 35(2):416–438, 2020.

[88] Baturalp Yalcin, Ziye Ma, Javad Lavaei, and Somayeh Sojoudi. Semidefinite programming versus burer-monteiro factorization for matrix sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[89] Peng Yi-Gang, Suo Jin-Li, DAI Qiong-Hai, and XU Wen-Li. From compressed sensing to low-rank matrix recovery: theory and applications. *Acta Automatica Sinica*, 39(7):981–994, 2013.

[90] Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On compressing deep models by low rank and sparse decomposition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7370–7379, 2017.

[91] Alp Yurtsever, Joel A. Tropp, Olivier Fercoq, Madeleine Udell, and Volkan Cevher. Scalable semidefinite programming. *SIAM Journal on Mathematics of Data Science*, 3(1):171–200, 2021.

[92] Alp Yurtsever, Madeleine Udell, Joel Tropp, and Volkan Cevher. Sketchy Decisions: Convex Low-Rank Matrix Optimization with Optimal Storage. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1188–1196. PMLR, 20–22 Apr 2017.

[93] Gavin Zhang and Richard Y. Zhang. How many samples is a good initial point worth in low-rank matrix recovery? In *Advances in Neural Information Processing Systems*, volume 33, pages 12583–12592, 2020.

[94] 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.

[95] Richard Y. Zhang. Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime, 2021. arXiv:2104.10790.

[96] Richard Y Zhang. Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime. *arXiv preprint arXiv:2104.10790*, 2021.

[97] Richard Y Zhang. Improved global guarantees for the nonconvex burer–monteiro factorization via rank overparameterization. *arXiv preprint arXiv:2207.01789*, 2022.

[98] Richard Y. Zhang, Cédric Josz, Somayeh Sojoudi, and Javad Lavaei. How much restricted isometry is needed in nonconvex matrix recovery? In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[99] Richard Y Zhang and Javad Lavaei. Sparse semidefinite programs with guaranteed near-linear time complexity via dualized clique tree conversion. *Mathematical programming*, 188(1):351–393, 2021.

[100] Richard Y Zhang, Javad Lavaei, and Ross Baldick. Spurious local minima in power system state estimation. *IEEE transactions on control of network systems*, 6(3):1086–1096, 2019.

[101] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 20(114):1–34, 2019.

[102] Xiao Zhang, Lingxiao Wang, Yaodong Yu, and Quanquan Gu. A primal-dual analysis of global optimality in nonconvex low-rank matrix recovery. In *International conference on machine learning*, pages 5862–5871. PMLR, 2018.

[103] Xiao Zhang, Lingxiao Wang, Yaodong Yu, and Quanquan Gu. A primal-dual analysis of global optimality in nonconvex low-rank matrix recovery. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5862–5871, 2018.

[104] Yu Zhang, Ramtin Madani, and Javad Lavaei. Conic relaxations for power system state estimation with line measurements. *IEEE Transactions on Control of Network Systems*, 5(3):1193–1205, 2017.

[105] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

[106] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B. Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

[107] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of low-rank matrix optimization. *IEEE Transactions on Information Theory*, 67(2):1308–1331, 2021.

[108] Jiacheng Zhuo, Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *arXiv preprint arXiv:2102.02756*, 2021.

[109] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109(3):467–492, 2020.