Absence of spurious solutions far from ground truth: A low-rank analysis with high-order losses

Ziye Ma[*,1] **Ying Chen**[*,2] **Javad Lavaei**[2] **Somayeh Sojoudi**[1]
Department of Electrical Engineering and Computer Science[1]
Department of Industrial Engineering and Operations Research[2]
University of California, Berkeley

Abstract

Matrix sensing problems exhibit pervasive non-convexity, plaguing optimization with a proliferation of suboptimal spurious solutions. Avoiding convergence to these critical points poses a major challenge. This work provides new theoretical insights that help demystify the intricacies of the non-convex landscape. In this work, 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.

1 Introduction

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 paper, we study this issue by focusing on a class of benchmark non-convex problems, named matrix sensing, and analyze the landscape of the optimization problem in areas far away from the ground truth.

To be more concrete, we focus on the Burer-Monteiro (BM) form of the problem:

$$\min_{X\in\mathbb{R}^{n\times r}} f(X) \coloneqq \frac{1}{2}\|\mathcal{A}(XX^T)-b\|^2 \tag{1}$$

where $b=\mathcal{A}(M^*)\in\mathbb{R}^m$ is the vector of observed measurements and $M^*\in\mathbb{R}^{n\times n}$ is the ground truth solution to be found. We focus on the case where M^* is positive semidefinite and symmetric since the asymmetric case (with M^* being sign indefinite or rectangular) can be equivalently converted to the symmetric case (Bi et al., 2022). The linear operator $\mathcal{A}(\cdot)$ in (1) is defined as $\mathcal{A}(M)=[\langle A_1,M\rangle,\dots,\langle A_m,M\rangle]^T$, where $\{A_i\}_{i=1}^m$ are m sensing matrices, which can be assumed to be symmetric without loss of generality (Zhang et al., 2021). We use r^* to denote the true rank of M^*. In (1), X can have a search rank of r, which has to satisfy $r\geq r^*$. Since r is often significantly smaller than n in practice, the above factorization form optimizing over the matrix X with nr entries rather than a matrix M with n^2 entries has major computational advantages.

The matrix sensing problem is a canonical problem bearing many important applications, such as the matrix completion problem/netflix problem (Candès and Recht, 2009; Candès and Tao, 2010), the compressed sensing problem (Donoho, 2006), the training

[*]These authors contributed equally to this work.
Preprint version.

of quadratic neural networks (Li et al., 2018), and an array of localization/estimation problems (Zhang et al., 2017; Jin et al., 2019; Singer, 2011; Boumal, 2016; Shechtman et al., 2015; Fattahi and Sojoudi, 2020). As a result, a better understanding of (1) 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 (Molybog et al., 2020).

The major drawback of (1) 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), and the centerpiece notion is the restricted isometry property (RIP), defined below.

Definition 1 (RIP)**.** (Candès and Recht, 2009) 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 $\text{rank}(M)\leq p$.

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 $b=\text{vec}(M^*)$, which makes the problem trivial to solve for M^*.

In this section, we provide a review of how RIP plays a central role in determining the optimization landscape of the non-convex problem (1), 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: RIP being smaller than 1/2 and RIP being greater than 1/2.

1.1 When RIP constant is smaller than 1/2

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 (Recht et al., 2010; Candès and Tao, 2010). It was proven that as along as $\delta_{5r^*}\leq 1/10$, the SDP relaxation was tight and M^* could be recovered exactly. Subsequently, Bhojanapalli et al. (2016) analyzed the factorized problem (1) and concluded that as long as $\delta_{2r}\leq 1/5$, all second-order critical points (SOPs) of (1) are ground truth solutions. Zhu et al. (2018); Li et al. (2019) 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)). Later, by using a "certification of in-existence" technique, Zhang et al. (2019) established that $\delta_{2r}=1/2$ was a sharp bound when $r=r^*$, meaning that as long as $\delta_{2r}<1/2$, all problem instances of (1) 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. This aforementioned approach is important because it quantifies how restrictive RIP needs to be in order to ensure a benign landscape.

Following the above line of work, Ma et al. (2022) proved that the same RIP bound of 1/2 is also applicable in noisy cases, and Bi et al. (2022) showed that even for general low-rank optimization problems beyond matrix sensing, $\delta_{2r}<1/2$ is still a sharp bound for the global recoverability of M^*, highlighting the importance of the 1/2 bound.

Furthermore, when the RIP constant is small enough, various desirable properties hold, including fast convergence (Li et al., 2018; Wang et al., 2017) and spectral contraction (Stöger and Soltanolkotabi, 2021; Jin et al., 2023).

1.2 When RIP constant is larger than 1/2

As proven in Zhang et al. (2021), when the RIP constant of the problem is larger or equal to 1/2, counterexamples can be found for which some SOPs are not global solutions. This means that in the regime where $\delta\geq 1/2$, the optimization landscape of (1) becomes complex. Several works have attempted to provide limited mathematical guarantees in that regime.

Benign landscape near M^*. Zhang et al. (2019) proved that when $\delta_{2r}\geq 1/2$ for $r=1$, we can ensure the absence of spurious solutions in a local region that is close to M^*, depending on the RIP constant and also the size of M^*. Zhang and Zhang (2020) expanded that analysis to the regime of general r. Subsequently, Ma and Sojoudi (2023) extended the result to noisy and general objectives, proving the ubiquity of this phenomenon.

Over-parametrization with $r\geq r^*$. This line of work is concerned with the case where the search rank r is greater than the true rank, which means that the complexity of the algorithm is increased. Zhang (2022) proved that if $r>r^*[(1+\delta_n)/(1-\delta_n)-1]^2/4$, with $r^*\leq r<n$, then every SOP $\hat{X}$ satisfies that $\hat{X}\hat{X}^\top=M^*$. Ma and Fattahi (2022) derived a similar result for l_1 loss under an RIP-type condition.

The SDP approach. This approach uses the conventional technique of convex relaxations to solve the matrix sensing problem (Recht et al., 2010; Candes and Plan, 2011). It was recently proven in Yalcin et al.

(2023) that as long as the RIP constant δ_{2r^*} is lower than the maximum of $1/2$ and $2r^*/(n+(n-2r^*)(2l-5))$, the global solution of the SDP relaxation corresponds to M^*. Since this bound approaches 1 as r^* increases, it is more appealing than using the factorized version (1) when the RIP constant is not small.

Lifting into tensor space. Recently, Ma et al. (2023b) proposed to lift the matrix decision variables of (1) into tensors and then optimize over tensors. This approach is inspired by the Sum-of-Squares hierarchies and the authors proved that spurious solutions will be converted into saddle points through lifting, further alleviating the highly non-convex landscape of (1) when δ_{2r} is close to 1. However, the shortcomings of this approach is that the complexity of the problem will increase exponentially if a high-order lifting is required.

Overall, although various studies have been conducted to address the optimization landscape of (1) when the RIP constant is larger than $1/2$, they either require to increase the complexity of the algorithm by a large margin (via over-parametrization $r \gg r^*$, SDP relaxation, or tensor optimization) or require to initialize the algorithm close to M^*. Therefore, the following question arises: *Does there exist meaningful global guarantees for* (1) *in the case of* $\delta \geq 1/2$ *without increasing the computational complexity of the problem drastically?* . In this paper, we offer a partial affirmative answer to this question via our notion of high-order losses.

1.3 Main Contributions

1. We prove that all critical points of (1) are strict saddles when reasonably away from M^*, with the intensity of the smallest eigenvalue of the Hessian at each saddle point being proportional to its distance to M^*. This result implies that there are no spurious solutions far away from M^*, and that it is possible to reach a vicinity of M^* with saddle-escaping algorithms even with poor initializations.

2. As a by-product of the above result, we derive sufficient conditions on M^* to ensure that there are no spurious solutions in the entire space even in the regime of high RIP constants.

3. We introduce the notion of high-order losses where a penalization term with a controllable degree is added to the objective function of (1) with the property that the penalty is zero at the ground truth. We show that critical points far away from M^* will still remain strict saddles, while the spurious solutions will be easier to escape as the degree of the loss increases. In other words, the landscape of the optimization problem is reshaped favorably by the inclusion of such penalties. Our result implies that increasing the complexity of the objective function serves as a viable alternative to increasing the complexity of the problem via over-parametrization.

1.4 Notations

The notation $M \succeq 0$ means that M is a symmetric and positive semidefinite (PSD) matrix. $\sigma_i(M)$ denotes the i-th largest singular value of a matrix M, and $\lambda_i(M)$ denotes the i-th largest eigenvalue of M. $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ respectively denotes the minimum and maximum eigenvalues of M. $\|v\|$ denotes the Euclidean norm of a vector v, while $\|M\|_F$ and $\|M\|_p$ denote the Frobenius norm and induced l_p norm of a matrix M, respectively for $p \geq 2$. $\langle A, B\rangle$ is defined to be $\text{trace}(A^T B)$ for two matrices A and B of the same size. For a matrix M, $\text{vec}(M)$ is the usual vectorization operation by stacking the columns of the matrix M into a vector. $[n]$ denotes the integer set $\{1, \ldots, n\}$.

The Hessian of the function $f(X)$ in (1), denoted as $\nabla^2 f(X) : \mathbb{R}^{n\times r} \times \mathbb{R}^{n\times r} \mapsto \mathbb{R}$, can be regarded as a quadratic form whose action on any two matrices $U, V \in \mathbb{R}^{n\times r}$ is given by

$$\nabla^2 f(X)[U,V] = \sum_{i,j,k,l=1}^{n,r,n,r} \frac{\partial^2}{\partial X_{ij}\partial X_{kl}} f(X) U_{ij} V_{kl}.$$

2 Disappearance of Spurious Solutions Far from Ground Truth

As discussed in Section 1, the optimization landscape of (1) 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. Before diving into more details, we present the first and second order critical conditions of (1):

Lemma 1. *A point X is a first-order critical point of* (1) *if*

$$\nabla f(X) = \left(\sum_{i=1}^{m} \langle A_i, XX^\top - M^*\rangle A_i\right) X = 0 \qquad (2)$$

and it is a second-order critical point if it satisfies the above condition together with

$$\begin{aligned} \nabla^2 f(X)[U,U] = \sum_{i=1}^{m} \langle A_i, UX^\top + XU^\top\rangle^2 + \\ \langle A_i, XX^\top - M^*\rangle \langle A_i, 2UU^\top\rangle \geq 0 \quad \forall U \in \mathbb{R}^{n\times r} \end{aligned} \qquad (3)$$

The proof of this lemma is plain calculus thus omitted for simplicity. Focusing on (3), it is apparent that for a first-order critical point $\hat{X}$ satisfying $\nabla f(\hat{X}) = 0$, the Hessian $\nabla^2 f(\hat{X})[U, U]$ can be broken down into the summation of two terms:

$$T_1 \coloneqq \sum_{i=1}^{m} \langle A_i, U\hat{X}^\top + \hat{X}U^\top \rangle^2 = \|\mathcal{A}(U\hat{X}^\top + \hat{X}U^\top)\|_2^2,$$

$$T_2 \coloneqq \sum_{i=1}^{m} \langle A_i, \hat{X}\hat{X}^\top - M^* \rangle \langle A_i, 2UU^\top \rangle$$
$$= 2\langle \nabla h(\hat{X}\hat{X}^\top), UU^\top \rangle$$

Assuming that the problem (1) 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 $U \in \mathbb{R}^{n\times r}$ 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

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

where $h(\cdot)$ is defined as

$$h(M) = \frac{1}{2}\|\mathcal{A}(M - M^*)\|^2 \tag{4}$$

Given $\nabla f(\hat{X}) = 0$ and the expression in (2), we obtain that

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

since $h(\hat{X}\hat{X}^\top) \geq h(M^*)$ 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 1, serving as the cornerstone of all results in this paper. A detailed proof can be found in the Appendix.

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

$$\|\hat{X}\hat{X}^\top - M^*\|_F^2 > 2\frac{1 + \delta}{1 - \delta}\operatorname{trace}(M^*)\sigma_r(\hat{X})^2, \tag{5}$$

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

$$2(1 + \delta)\sigma_r(\hat{X})^2 - \frac{\|\hat{X}\hat{X}^\top - M^*\|_F^2(1 - \delta)}{\operatorname{trace}(M^*)} \tag{6}$$

Theorem 1 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) is poorly initialized, it is possible to converge to a vicinity of M^* with saddle-escaping algorithms, which we will numerically illustrate in Section 4.

In contrary to Theorem 1, the existing results in the literature state that there are no spurious solutions in a small neighborhood of M^* (Zhang and Zhang, 2020; Bi and Lavaei, 2020; Ma et al., 2023a). We recall a classic result in the literature regarding this property below.

Theorem 2 (Zhang and Zhang (2020))**.** *Assume that* (1) *satisfies the RIP property with constant $\delta \in [0, 1)$. Given an arbitrary constant $\tau \in (0, 1 - \delta^2)$, if a second-order critical point $\hat{X} \in \mathbb{R}^{n\times r}$ of* (1) *satisfies*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \tau\lambda_{r^*}(M^*) \tag{7}$$

then $\hat{X}$ corresponds to the ground truth solution.

A question arises as to whether there is a sufficient condition to ensure that the two regions (5) and (7) overlap? An affirmative answer guarantees that the entire optimization landscape has no spurious solutions, even if $\delta > 1/2$, exceeding the sharp RIP bound regardless of $\mathcal{A}$ and M^* Zhang et al. (2019). In what follows, we will prove that if $\|M^*\|_F$ is small, the regions (5) and (7) overlap.

Theorem 3. *Consider the problem* (1) *under the RIP$_{r+r^*}$ property with a constant $\delta \in [0, 1)$. Assume that its ground truth solution M^* satisfies the following inequality*

$$\|M^*\|_F \frac{\operatorname{trace}(M^*)}{\lambda_{r^*}(M^*)} \leq \frac{\sqrt{r}}{2\sqrt{2}}\sqrt{\frac{(1 - \delta)^5}{(1 + \delta)}}, \tag{8}$$

Then, every second-order critical point $\hat{X}$ of (1) *satisfies*

$$\hat{X}\hat{X}^\top = M^*$$

The proof can be found in the Appendix.

3 Higher-order Loss Functions

Although Theorem 1 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 (Ge et al., 2017; Jin et al., 2021). Therefore, it is essential to understand whether the curvatures at saddle points could be enhanced to reshape the landscape favorably. In this section, we

n	λ	$\lambda_{\min}(\nabla^2 f^l(\hat{X}))$	$\lambda_{\max}(\nabla^2 f^l(\hat{X}))$	$\lambda_{\min}(\nabla^2 f^l(X^*))$	$\lambda_{\max}(\nabla^2 f^l(X^*))$
3	0	1.821	3.642	2.18	4.36
3	0.5	1.779	3.855	2.18	4.36
3	5	1.594	7.422	2.18	4.36
3	50	1.470	55.028	2.18	4.36
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	0.900	5.440
9	0.5	0.601	4.315	0.900	5.440
9	5	0.557	10.915	0.900	5.440
9	50	0.528	84.002	0.900	5.440

Table 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 1 shows that there are no such spurious solutions). The problem satisfies the RIP_{2r}-property with $\delta = \frac{1-\varepsilon}{1+\varepsilon} = 0.538 > 1/2$, and hence has spurious local minima.

provide an affirmative answer to this question by using a modified loss function[1].

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) to achieve this goal. An l_2 loss function is used in (1) 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:

$$\min_{X \in \mathbb{R}^{n \times r}} f_\lambda^l(X) \coloneqq f(X) + \lambda f^l(X) \tag{9}$$

where

$$f^l(X) \coloneqq \frac{1}{l} \|\mathcal{A}(XX^\top) - b\|_l^l \tag{10a}$$

$$h^l(M) \coloneqq \frac{1}{l} \|\mathcal{A}(M) - b\|_l^l \tag{10b}$$

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:

$$\min_{x \in \mathbb{R}} g(x) \coloneqq \frac{1}{l}(x^2 - a)^l \tag{11}$$

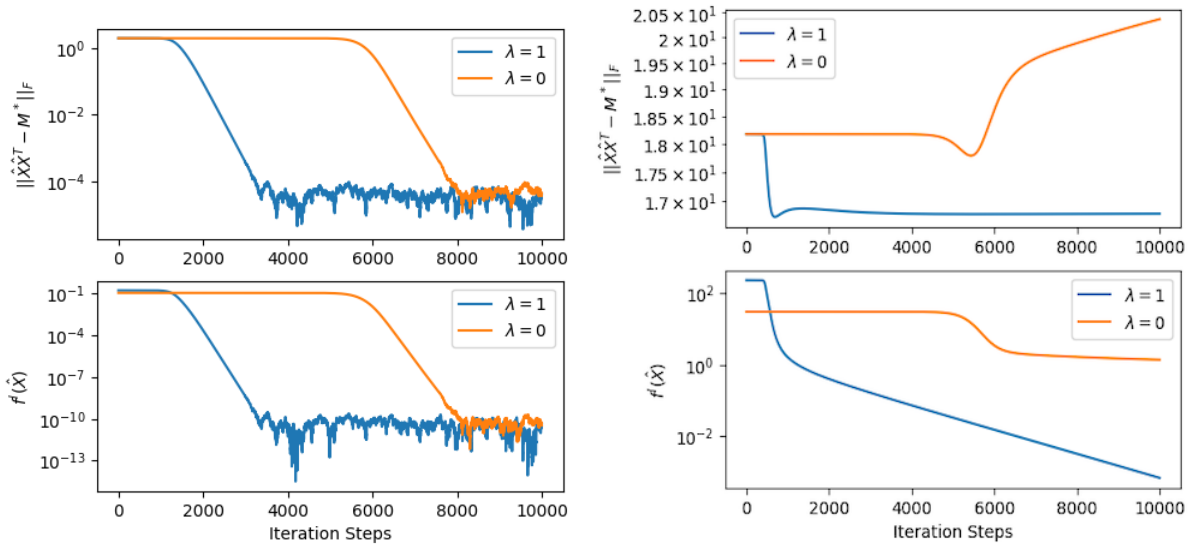
for some constant $a \in \mathbb{R}$ and an even number $l \geq 2$. This problem is a scalar analogy of $f^l(X)$ with $\mathcal{A}(\cdot)$ being the identity operator. In this example, the derivatives are

$$g'(x) = 2x(x^2 - a)^{l-1},$$
$$g''(x) = 2(x^2 - a)^{l-2}\left[(l-1)2x^2 + (x^2 - a)\right]$$

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 $f_\lambda^l(X)$ instead of $f^l(X)$. If we directly minimize $f^l(X)$ 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 (Wright and Recht, 2022). Conversely, when using the original objective (1), we see from (3) that even if $XX^\top$ is close to M^*, the Hessian is positive semidefinite, and therefore adding $f^l(X)$ to the objective of (1) 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 $f^l(X)$ (see Lemma 2 in Appendix) is very small due to its exponentiation nature. This means that minimizing $f^l(X)$ alone will suffer from the vanishing

[1]The code to this section can be found at https://github.com/anonpapersbm/high_order_obj



(a) $\lambda = 0$ converges to ground truth

(b) $\lambda = 0$ converges to a spurious solution around the ground truth

Figure 1: 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.

issue and slow growth rate in a local region around M^*.

Due to the above reasons, we mix $f^l(X)$ with the original objective in (1) 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 (9), we can arrive at a similar result to Theorem 1

Theorem 4. *Assume that the operator $\mathcal{A}(\cdot)$ satisfies the RIP$_{r+r^*}$ property with constant $\delta \in [0, 1)$. Consider the high-order optimization problem (9) such that $l \geq 2$ is even. Given a first-order critical point $\hat{X} \in \mathbb{R}^{n\times r}$ of (9), if*

$$D^2 \geq \operatorname{trace}(M^*)\sigma_r^2(\hat{X})\frac{(1+\delta)+\lambda(l-1)(1+\delta)^{l/2}D^{l-2}}{(1-\delta)/2+\lambda C(l)(1-\delta)^{l/2}D^{l-2}}, \tag{12}$$

then $\hat{X}$ is a strict saddle point with $\nabla^2 f(\hat{X})$ having a strictly negative eigenvalue not larger than

$$\left[2(1+\delta)\sigma_r(\hat{X})^2 - \frac{D^2(1-\delta)}{\operatorname{trace}(M^*)}\right] + \lambda D^{l-2}\left[2(1+\delta)^{l/2}(l-1)\sigma_r(\hat{X})^2 - 2\frac{(1-\delta)^{l/2}C(l)D^2}{\operatorname{trace}(M^*)}\right] \tag{13}$$

where

$$D \coloneqq \|\hat{X}\hat{X}^\top - M^*\|_F, \qquad C(l) \coloneqq m^{(2-l)/2}\left(\frac{2^l-1}{l} - 1\right) \tag{14}$$

Theorem 4 serves as a direct generalization of Theorem 1, as it recovers the statements of Theorem 1 when l is set to 2 or λ is set to 0. By comparing (13) to (6), 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 (9) compared to (1) when $\|\hat{X}\hat{X}^\top - M^*\|_F$ is large.

Theorem 4 contrasts well with an existing approach that utilizes a lifting technique to eliminate spurious solutions. Ma et al. (2023b) 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 4 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 (Ma et al., 2023b) can amplify the negative curvature of those points X that satisfy

$$\|XX^\top - M^*\|_F^2 \geq \frac{1+\delta}{1-\delta}\operatorname{trace}(M^*)\sigma_r^2(\hat{X})$$

where in comparison to (12) the multiplicative factor to $\operatorname{trace}(M^*)\sigma_r^2(\hat{X})$ 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

$$\mathcal{O}\left(l\left(\frac{\sqrt{m}}{2}\right)^{l}\left(\frac{1+\delta}{1-\delta}\right)^{l/2}\right),$$

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.

4 Simulation 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). 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

$$\mathcal{A}_\epsilon(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij}, & \text{if } (i,j) \in \Omega \\ \epsilon \mathbf{M}_{ij}, & \text{otherwise} \end{cases}, \tag{15}$$

where $\Omega = \{(i,i),(i,2k),(2k,i) \mid \forall i \in [n], k \in [\lfloor n/2 \rfloor]\}$, $0 < \epsilon < 1$. Yalcin et al. (2023) has proved that while satisfying RIP property with $\delta_{2r} = (1-\epsilon)/(1+\epsilon)$, 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) and the altered problem (9) 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 1). The findings of this study are presented in Table 1, while the ratio between the largest and smallest eigenvalues at the spurious local minimum $\hat{X}$ is plotted in Figure 2.

Table 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

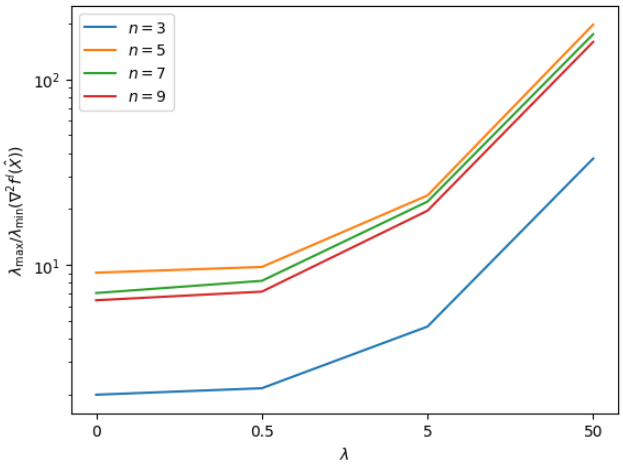


Figure 2: The ratio between the largest and smallest eigenvalue of Hessian at the spurious local minimum $\lambda_{\max}/\lambda_{\min}(\nabla^2 f^l(\hat{X}))$ with respect to λ under different size n.

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 1, 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 $f^l(\hat{X})$. Although Figure 1 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.

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

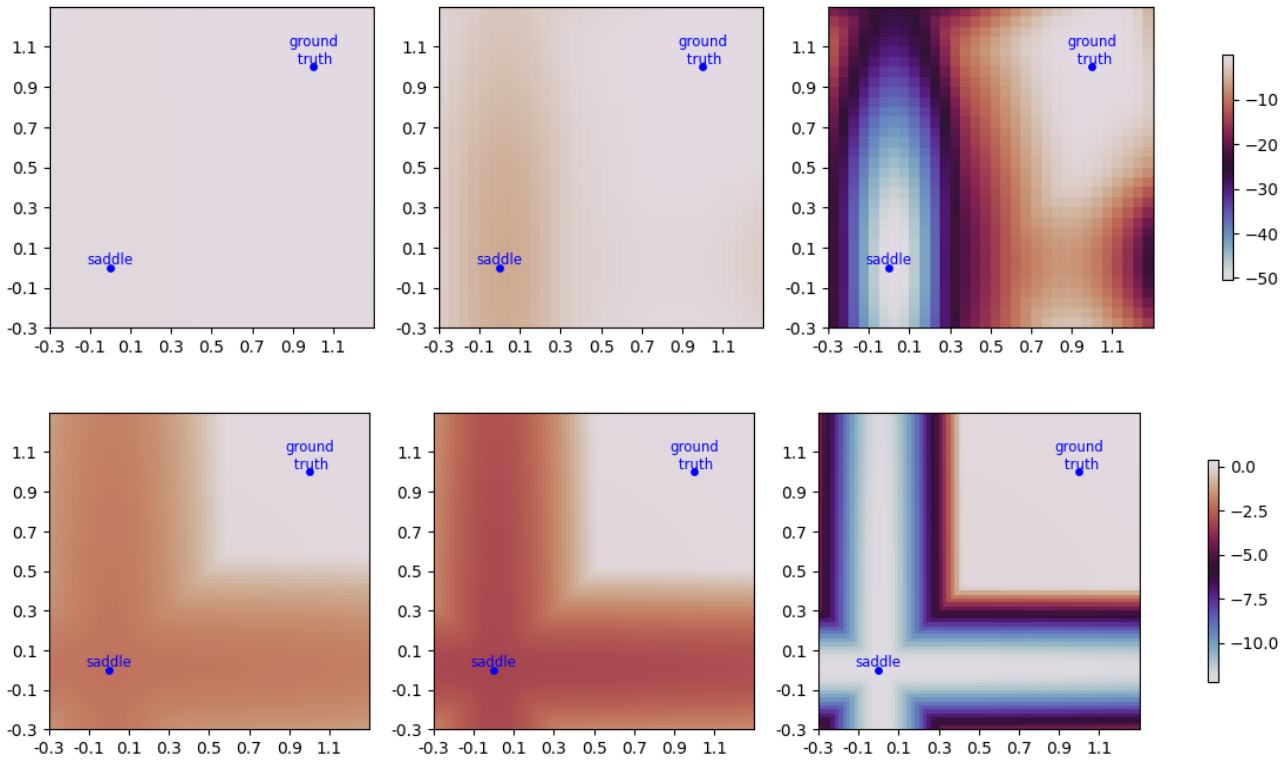


Figure 3: The value of the minimum eigenvalue of the Hessian around saddle points: The first row is for randomly generated Gaussian matrix with $m = 20, n = 20$, and the second row is for problem (15) with $n = 21, \epsilon = 0.1$. $\lambda = 0$ (left column), $\lambda = 0.5$ (middle column), $\lambda = 5$ (right column), with x-axis and y-axis as two orthogonal directions from the critical point to the ground truth.

size $m = 20, n = 20$, and the problem (15) with the parameters $n = 21$ and $l = 4$. The result is plotted in Figure 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 4 (Thereby Theorem 1). For instance, when $\lambda = 0.5$, the minimum eigenvalue of Hessian matrix at the first-order critical point $\hat{X}$ is $\lambda_{\min}(\nabla^2 f^l(\hat{X})) = -3.201$, which is smaller than the eigenvalue-bound -2.274; the distance from the ground truth matrix is $D \coloneqq \|\hat{X}\hat{X}^\top - M^*\|_F = 11.0$, larger than the distance bound in (12), validating Theorem 4.

5 Conclusion

This work 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.

References

Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016). Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 29.

Bi, Y. and Lavaei, J. (2020). Global and local analyses of nonlinear low-rank matrix recovery problems. arXiv:2010.04349.

Bi, Y., Zhang, H., and Lavaei, J. (2022). Local and global linear convergence of general low-rank matrix recovery problems. *AAAI-22*.

Boumal, N. (2016). Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377.

Candes, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

Fattahi, S. and Sojoudi, S. (2020). 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.

Ge, R., Jin, C., and Zheng, Y. (2017). 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.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2021). On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29.

Jin, J., Li, Z., Lyu, K., Du, S. S., and Lee, J. D. (2023). Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. *arXiv preprint arXiv:2301.11500*.

Jin, M., Molybog, I., Mohammadi-Ghazi, R., and Lavaei, J. (2019). Towards robust and scalable power system state estimation. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3245–3252. IEEE.

Li, Q., Zhu, Z., and Tang, G. (2019). The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96.

Li, Y., Ma, T., and Zhang, H. (2018). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR.

Ma, J. and Fattahi, S. (2022). Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *arXiv preprint arXiv:2202.08788*.

Ma, Z., Bi, Y., Lavaei, J., and Sojoudi, S. (2022). 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.

Ma, Z., Bi, Y., Lavaei, J., and Sojoudi, S. (2023a). Geometric analysis of noisy low-rank matrix recovery in the exact parametrized and the overparametrized regimes. *INFORMS Journal on Optimization*.

Ma, Z., Molybog, I., Lavaei, J., and Sojoudi, S. (2023b). Over-parametrization via lifting for low-rank matrix sensing: Conversion of spurious solutions to strict saddle points. In *International Conference on Machine Learning*. PMLR.

Ma, Z. and Sojoudi, S. (2023). Noisy low-rank matrix optimization: Geometry of local minima and convergence rate. In *International Conference on Artificial Intelligence and Statistics*, pages 3125–3150. PMLR.

Molybog, I., Madani, R., and Lavaei, J. (2020). Conic optimization for quadratic regression under sparse noise. *The Journal of Machine Learning Research*, 21(1):7994–8029.

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501.

Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. (2015). Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109.

Singer, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36.

Stöger, D. and Soltanolkotabi, M. (2021). 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.

Wang, L., Zhang, X., and Gu, Q. (2017). 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.

Wright, S. J. and Recht, B. (2022). *Optimization for data analysis*. Cambridge University Press.

Yalcin, B., Ma, Z., Lavaei, J., and Sojoudi, S. (2023). Semidefinite programming versus burer-monteiro factorization for matrix sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, G. and Zhang, R. Y. (2020). 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.

Zhang, H., Bi, Y., and Lavaei, J. (2021). General low-rank matrix optimization: Geometric analysis and sharper bounds. *Advances in Neural Information Processing Systems*, 34:27369–27380.

Zhang, R. Y. (2022). Improved global guarantees for the nonconvex burer–monteiro factorization via rank overparameterization. *arXiv preprint arXiv:2207.01789*.

Zhang, R. Y., Sojoudi, S., and Lavaei, J. (2019). Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 20(114):1–34.

Zhang, Y., Madani, R., and Lavaei, J. (2017). Conic relaxations for power system state estimation with line measurements. *IEEE Transactions on Control of Network Systems*, 5(3):1193–1205.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. (2018). Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628.

A Appendix

A.1 Optimality Conditions

Lemma 2. *Given the problem* (9)*, its gradient and Hessian are given as*

$$\langle \nabla f^l(X), U\rangle = \sum_{i=1}^m \langle A_i, XX^\top - M^*\rangle^{l-1}\langle A_i, UX^\top + XU^\top\rangle \quad \forall U \in \mathbb{R}^{n\times r}, \tag{16a}$$

$$\nabla^2 f^l(X)[U,U] = \sum_{i=1}^m \langle A_i, XX^\top - M^*\rangle^{l-2}\left[(l-1)\langle A_i, UX^\top + XU^\top\rangle^2 + \langle A_i, XX^\top - M^*\rangle\langle A_i, 2UU^\top\rangle\right] \quad \forall U \in \mathbb{R}^{n\times r} \tag{16b}$$

Lemma 3. *Given the problem* (10b)*, its gradient, Hessian and high-order derivatives are equal to*

$$\nabla h^l(M) = \sum_{i=1}^m \langle A_i, M - M^*\rangle^{l-1} A_i, \tag{17a}$$

$$\begin{aligned}
&\nabla^2 h^l(M)[N,N] = (l-1)\sum_{i=1}^m \langle A_i, M - M^*\rangle^{l-2}\langle A_i, N\rangle^2 \quad \forall N \in \mathbb{R}^{n\times n},\\
&\nabla^p h^l(M)[\underbrace{N,\ldots,N}_{p \text{ times}}] = \frac{(l-1)!}{(l-p)!}\sum_{i=1}^m \langle A_i, M - M^*\rangle^{l-p}\langle A_i, N\rangle^p \quad \forall N \in \mathbb{R}^{n\times n}
\end{aligned} \tag{17b}$$

A.2 Proofs in Section 2

Proof of Theorem 1. Via the definition of RIP, we have that

$$h(M^*) \geq h(\hat{X}\hat{X}^\top) + \langle \nabla h(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top\rangle + \frac{1-\delta}{2}\|\hat{X}\hat{X}^\top - M^*\|_F^2$$

Given that $\hat{X}$ is a first-order critical point, it means that it must satisfy first-order optimality conditions, which means

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

leading to

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

since $h(\hat{X}\hat{X}^\top) - h(M^*) \geq 0$ 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 (Zhang et al., 2021), and the fact that M^* is positive-semidefinite, we have that

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

This leads to the fact that

$$\lambda_{\min}(\nabla h(\hat{X}\hat{X}^\top)) \leq -\frac{(1-\delta)\|\hat{X}\hat{X}^\top - M^*\|_F^2}{2\,\mathrm{trace}(M^*)} \leq 0 \tag{18}$$

Given the optimality conditions, we know that for $\hat{X}$ to be a strict saddle, we must prove that there exists a direction $\Delta \in \mathbb{R}^{n\times r}$ such that

$$2\langle \nabla h(\hat{X}\hat{X}^\top), \Delta\Delta^\top\rangle + \underbrace{\left[\nabla^2 h(\hat{X}\hat{X}^\top)\right](\hat{X}\Delta^\top + \Delta\hat{X}^\top, \hat{X}\Delta^\top + \Delta\hat{X}^\top)}_{P(\Delta)} < 0 \tag{19}$$

If we choose

$$\Delta = uq^\top, \quad \|u\|_2, \|q\|_2 = 1, \quad \|\hat{X}q\|_2 = \sigma_r(\hat{X}),\ u^\top\nabla h(\hat{X}\hat{X}^\top)u = \lambda_{\min}(\nabla h(\hat{X}\hat{X}^\top)),$$

then it follows from the RIP condition that

$$\begin{aligned} P(\Delta) &\le (1+\delta)\|\hat{X}\Delta^\top + \Delta\hat{X}^\top\|_F^2 \\ &= (1+\delta)\|u(\hat{X}q)^\top + (\hat{X}q)u^\top\|_F^2 \\ &= 2(1+\delta)\|\hat{X}q\|_F^2 + 2(1+\delta)\left(q^\top(\hat{X}^\top u)\right)^2 \\ &= 2(1+\delta)\sigma_r(\hat{X})^2 \end{aligned}$$

because of $\hat{X}^\top u = 0$ due to the first-order optimality condition. Therefore,

$$\begin{aligned} 2\langle\nabla h(\hat{X}\hat{X}^\top), \Delta\Delta^\top\rangle + P(\Delta) &\le 2\langle h(\hat{X}\hat{X}^\top), \Delta\Delta^\top\rangle + 2(1+\delta)\sigma_r(\hat{X})^2 \\ &= 2\langle\nabla h(\hat{X}\hat{X}^\top), uu^\top\rangle + 2(1+\delta)\sigma_r(\hat{X})^2 \\ &= 2u^\top\nabla h(\hat{X}\hat{X}^\top)u + 2(1+\delta)\sigma_r(\hat{X})^2 \\ &= 2\left(\lambda_{\min}(\nabla h(\hat{X}\hat{X}^\top)) + (1+\delta)\sigma_r(\hat{X})^2\right) \\ &\le 2(1+\delta)\sigma_r(\hat{X})^2 - \frac{(1-\delta)\|\hat{X}\hat{X}^\top - M^*\|_F^2}{\operatorname{trace}(M^*)} \end{aligned}$$

Therefore, in order to make (19) hold, we simply need

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

which concludes the proof. □

Proof for Theorem 3. An easy extension of Lemma B.2 from Ma et al. (2023b) gives us that

$$\sigma_r^2(\hat{X}) < \sqrt{\frac{2(1+\delta)}{r(1-\delta)}}\|M^*\|_F \tag{20}$$

and we omit the proof for brevity. Then the only remaining step is to show that the right-hand side of (7) is larger than that of (5). This means the following equation

$$(1-\delta^2)\lambda_{r^*}(M^*) \ge 2\frac{1+\delta}{1-\delta}\operatorname{trace}(M^*)\sqrt{\frac{2(1+\delta)}{r(1-\delta)}}\|M^*\|_F$$

is sufficient to

$$(1-\delta^2)\lambda_{r^*}(M^*) \ge 2\frac{1+\delta}{1-\delta}\operatorname{trace}(M^*)\sigma_r^2(\hat{X})$$

which yields (8) after rearrangements. □

A.3 Proofs in Section 3

Before proceeding to the main proof, we first present a technical lemma:

Lemma 4. *Given a vector $x \in \mathbb{R}^n$, we have that*

$$\|x\|_p \le n^{1/p-1/q}\|x\|_q \quad \forall\ q \ge p \tag{21}$$

Proof of Lemma 4. By applying Holder's inequality, we obtain that

$$\sum_{i=1}^n |x_i|^p = \sum_{i=1}^n |x_i|^p \cdot 1 \le \left(\sum_{i=1}^n (|x_i|^p)^{\frac{q}{p}}\right)^{\frac{p}{q}} \left(\sum_{i=1}^n 1^{\frac{q}{q-p}}\right)^{1-\frac{p}{q}} = \left(\sum_{i=1}^n |x_i|^q\right)^{\frac{p}{q}} n^{1-\frac{p}{q}}$$

Then

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \le \left(\left(\sum_{i=1}^n |x_i|^q\right)^{\frac{p}{q}} n^{1-\frac{p}{q}}\right)^{1/p} = \left(\sum_{i=1}^n |x_i|^q\right)^{\frac{1}{q}} n^{\frac{1}{p}-\frac{1}{q}} = n^{1/p-1/q}\|x\|_q$$

□

Proof to Theorem 4. First, we define

$$h_\lambda^l(M) = h(M) + \lambda h^l(M)$$

Then, focusing on $h^l(\cdot)$, using Taylor's theorem and the mean-value theorem, we have

$$\begin{aligned} h^l(M^*) =& h^l(\hat{X}\hat{X}^\top) + \langle \nabla h^l(\hat{X}\hat{X}^\top), \Delta\rangle + \frac{1}{2!}\nabla^2 h^l(\hat{X}\hat{X}^\top)[\Delta,\Delta] + \\ & \frac{1}{3!}\nabla^3 h^l(\hat{X}\hat{X}^\top)[\Delta,\Delta,\Delta] + \cdots + \frac{1}{l!}\nabla^l h^l(\tilde{M})[\Delta,\ldots,\Delta] \end{aligned} \tag{22}$$

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 3, we know that

$$\begin{aligned} \frac{1}{p!}\nabla^p h^l(\hat{X}\hat{X}^\top)[\Delta,\ldots,\Delta] &= \frac{1}{p!}\sum_{i=1}^m \langle A_i, \Delta\rangle^l \quad \forall p \in [2, l-1], \\ \frac{1}{l!}\nabla^l h^l(M)[\Delta,\ldots,\Delta] &= \frac{1}{l!}\sum_{i=1}^m \langle A_i, \Delta\rangle^l \quad \forall M \in \mathbb{R}^{n\times n} \end{aligned}$$

Hence, it is possible to rewrite

$$\begin{aligned} h^l(M^*) &= h^l(\hat{X}\hat{X}^\top) + \langle \nabla h^l(\hat{X}\hat{X}^\top), \Delta\rangle + \sum_{p=2}^{l} \frac{(l-1)!}{(l-p)!p!}\sum_{i=1}^m \langle A_i, \Delta\rangle^l \\ &= h^l(\hat{X}\hat{X}^\top) + \langle \nabla h^l(\hat{X}\hat{X}^\top), \Delta\rangle + \sum_{i=1}^m \langle A_i, \Delta\rangle^l \sum_{p=2}^{l} \frac{(l-1)!}{(l-p)!p!} \\ &= h^l(\hat{X}\hat{X}^\top) + \langle \nabla h^l(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top\rangle + (\frac{2^l-1}{l} - 1)\|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_l^l \end{aligned} \tag{23}$$

By Lemma 4, 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

$$\|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_l^l \geq (1-\delta)^{l/2}\|\hat{X}\hat{X}^\top - M^*\|_F^l m^{(2-l)/2} \tag{24}$$

Therefore we know that

$$h_\lambda^l(M^*) = h(M^*) + \lambda h^l(M^*) \geq h_\lambda^l(\hat{X}\hat{X}^\top) + \langle \nabla h_\lambda^l(\hat{X}\hat{X}^\top), M^* - \hat{X}\hat{X}^\top\rangle + L \tag{25}$$

in which

$$L \coloneqq \frac{1-\delta}{2}\|M^* - \hat{X}\hat{X}^\top\|_F^2 + \lambda(1-\delta)^{l/2}C(l)\|M^* - \hat{X}\hat{X}^\top\|_F^l$$

If $\hat{X}$ is a first-order critical point, a repeated application of (16a) yields that

$$\nabla h_\lambda^l(\hat{X}\hat{X}^\top)\hat{X} = 0 \implies \langle \nabla h_\lambda^l(\hat{X}\hat{X}^\top), \hat{X}\hat{X}^\top\rangle = 0$$

which after rearrangement leads to

$$\begin{aligned} \langle \nabla h_\lambda^l(\hat{X}\hat{X}^\top), M^*\rangle &\leq \left[h_\lambda^l(M^*) - h_\lambda^l(\hat{X}\hat{X}^\top)\right] - L \\ &\leq -L \end{aligned} \tag{26}$$

where the second inequality follows from the fact that M^* is the global minimizer of (9). Since the sensing matrices can be assumed to be symmetric without loss of generality (Zhang et al., 2021), $\nabla h^l(\hat{X}\hat{X}^\top)$ can be assumed to be symmetric according to (17a). This means that

$$\langle \nabla h_\lambda^l(\hat{X}\hat{X}^\top), M^*\rangle \geq \operatorname{trace}(M^*)\lambda_{\min}(\nabla h_\lambda^l(\hat{X}\hat{X}^\top))$$

which further leads to

$$\lambda_{\min}(\nabla h_\lambda^l(\hat{X}\hat{X}^\top)) \leq -\frac{L}{\operatorname{trace}(M^*)} \tag{27}$$

Now, we turn to the Hessian of $f_\lambda^l(\cdot)$, which given (16b) is

$$\begin{aligned}\nabla^2 f_\lambda^l(\hat{X})[U,U] = &\underbrace{\sum_{i=1}^m \left[\lambda(l-1)\langle A_i, \hat{X}\hat{X}^\top - M^*\rangle^{l-2} + 1\right]\langle A_i, U\hat{X}^\top + \hat{X}U^\top\rangle^2}_{B} \\ &+ \underbrace{\sum_{i=1}^m \left[\lambda\langle A_i, \hat{X}\hat{X}^\top - M^*\rangle^{l-1} + \langle A_i, \hat{X}\hat{X}^\top - M^*\rangle\right]\langle A_i, 2UU^\top\rangle}_{A} \quad \forall U \in \mathbb{R}^{n\times r}\end{aligned} \tag{28}$$

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

$$U = uq^\top, \quad \|u\|_2, \|q\|_2 = 1, \quad \|\hat{X}q\|_2 = \sigma_r(\hat{X}),\ u^\top \nabla h^l(\hat{X}\hat{X}^\top)u = \lambda_{\min}(\nabla h_\lambda^l(\hat{X}\hat{X}^\top)),$$

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 $\nabla h_\lambda^l(\hat{X}\hat{X}^\top)$, which is orthogonal to $\hat{X}$ as required by (16a). This further implies that

$$\begin{aligned}B &\leq 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)\sum_{i=1}^m \langle A_i, \hat{X}\hat{X}^\top - M^*\rangle^{l-2} + 1\right] \\ &= 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)\|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_{l-2}^{l-2} + 1\right] \\ &\leq 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)(\|\mathcal{A}(\hat{X}\hat{X}^\top - M^*)\|_2)^{l-2} + 1\right] \\ &\leq 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)(\sqrt{1+\delta}\|\hat{X}\hat{X}^\top - M^*\|_F)^{l-2} + 1\right] \\ &= 2(1+\delta)\sigma_r(\hat{X})^2\left[\lambda(l-1)(1+\delta)^{(l-2)/2}\|\hat{X}\hat{X}^\top - M^*\|_F^{l-2} + 1\right]\end{aligned} \tag{29}$$

Also, given (17a) and our choice of U, it is apparent that

$$A = 2u^\top \nabla h_\lambda^l(\hat{X}\hat{X}^\top)u = 2\lambda_{\min}(\nabla h_\lambda^l(\hat{X}\hat{X}^\top)) \tag{30}$$

Therefore, given (29), (30) and (27), by substituting them back into (28), we arrive at

$$\nabla^2 f^l(\hat{X})[U,U] \leq 2\sigma_r(\hat{X})^2\left(\lambda(l-1)(1+\delta)^{l/2}\|\hat{X}\hat{X}^\top - M^*\|_F^{l-2} + (1+\delta)\right) - 2L/\operatorname{trace}(M^*) \tag{31}$$

so the right-hand side of the above equation is strictly negative if (12) is satisfied. □