Over-parametrization via Lifting for Low-rank Matrix Sensing: Conversion of Spurious Solutions to Strict Saddle Points

Ziye Ma [1] **Igor Molybog** [2] **Javad Lavaei** [3] **Somayeh Sojoudi** [1]

Abstract

This paper studies the role of over-parametrization in solving non-convex optimization problems. The focus is on the important class of low-rank matrix sensing, where we propose an infinite hierarchy of non-convex problems via the lifting technique and the Burer-Monteiro factorization. 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. This is the first result in the literature showing that over-parametrization creates a negative curvature for escaping spurious solutions. We also derive a bound on how much over-parametrization is requited to enable the elimination of spurious solutions.

1. Introduction

In this paper, we focus on an important class of non-convex optimization problems, named matrix sensing, which can be formulated as the feasibility problem:

$$\begin{aligned} \text{find} \quad & M \in \mathbb{R}^{n\times n} \\ \text{s.t.} \quad & \mathcal{A}(M) = \mathcal{A}(M^*) \\ & \operatorname{rank}(M) \le r, M \succeq 0. \end{aligned} \tag{1}$$

where the measurement operator $\mathcal{A}(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ is a linear operator returning a d-dimensional measurement vector $\mathcal{A}(M) = [\langle A_1, M\rangle, \dots, \langle A_m, M\rangle]^T$, for given sensing matrices $\{A_i\}_{i=1}^m \in \mathbb{R}^{n\times n}$. The goal is to find an unknown matrix M^* of rank r associated with the measurement vector b, meaning that $b = \mathcal{A}(M^*)$, where r is often much smaller than n. We factorize M^* as $M^* = ZZ^\top$ where $Z \in \mathbb{R}^{n\times r}$.

The problem (1) is extremely broad since solving any arbitrary polynomial optimization can converted to a series of problems in the form of (1) (Molybog et al., 2020). In addition, the problem (1) directly arises in various applications such as collaborative filtering (Koren et al., 2009), phase retrieval (Singer, 2011; Boumal, 2016; Shechtman et al., 2015), motion detection (Fattahi & Sojoudi, 2020), and power system state estimation (Zhang et al., 2017; Jin et al., 2019). Moreover, its strikingly simple form is associated with only one source of non-convexity, which is the rank constraint. As a result, the existing works have extensively studied under what conditions one can recover M^* exactly, and the centerpiece of this line of research is the notion of Restricted Isometry Property(RIP) of the measuring operator $\mathcal{A}$, which we state below:

Definition 1.1 (RIP)**.** (Candès & 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

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

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

This criterion is also intuitive to understand, as it measures how close $\mathcal{A}$ is to an identity operator (isometry) over matrices of rank at most r. If $\mathcal{A}$ is an exact isometry, or equivalently when it satisfies RIP with $\delta_r = 0$, we measure M^* exactly and the recovery is trivial. Therefore, a small value for the RIP constant usually implies that the problem has a low computational complexity.

A popular approach to solving (1) is to use the so-called Burer-Monteiro (BM) factorization (Burer & Monteiro, 2003). The BM formulation explicitly factors M as $M = XX^\top$, where $X \in \mathbb{R}^{n\times r}$

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

The problem (2) is an unconstrained smooth optimization problem, which means that highly scalable local search

[1]Department of EECS, UC Berkeley, USA [2]Meta AI [3]Department of IEOR, UC Berkeley, USA. Correspondence to: Ziye Ma <ziyema@berkeley.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

methods such as Gradient descent can be utilized to numerically solve it. Since the search is over $\mathbb{R}^{n\times r}$ instead of $\mathbb{R}^{n\times n}$ in the original feasibility problem (1), the number of variables is dramatically reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(nr)$, thereby improving its scalability. The main issue with (2) is that it is still a non-convex problem and thus it may contain spurious[1] local minima, preventing local search methods from convergence to a global optimum. However, despite its non-convexity, a recent line of work (Zhang et al., 2019b; Bi & Lavaei, 2020; Ma et al., 2023; Zhang et al., 2021; Ma et al., 2022; Ma & Sojoudi, 2023) has shown that if (2) satisfies the RIP condition with $\delta_{2r} < 1/2$, every local minimizer $\hat{X}$ of (2) will satisfy the relation $\hat{X}\hat{X}^\top = M^*$, precisely in the noiseless scenario and approximately when b is corrupted by random noise. 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 (1) (Recht et al., 2010; Candes & Plan, 2011; Cai & Zhang, 2013).

1.1. The power and limitation of over-parametrization

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 (Zhang et al., 2019a). 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 (2) (Zhang, 2021; 2022; Levin et al., 2022), 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 $r_{\text{search}} \in [r, n)$, and we arrive at the following counterpart of (2)

$$\min_{X\in\mathbb{R}^{n\times r_{\text{search}}}} f(X) := \frac{1}{2}\|\mathcal{A}(XX^T) - b\|^2. \tag{3}$$

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 1.2 (Theorem 1.1 of Zhang (2022))**.** *Assume that* (3) *satisfies the* (L_s, n)*-RSS (Restricted Strong Smoothness) and* (α_s, n)*-RSC (Restricted Smooth Convexity) properties. If*

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

then every second-order point (SOP) $\hat{X} \in \mathbb{R}^{n\times r_{search}}$ *of* (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 (3) over (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 (4) 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.

Our approach is related to over-parametrization used in the Semidefinite Programming (SDP) formulation. The SDP formulation is a natural convex relaxation of the original problem (1), obtained by removing the rank constrained. It aims to minimize the nuclear norm of M as a surrogate of its rank. When the search is performed on the space of symmetric and positive-semidefinite matrices, we can further reformulate the problem using a trace objective instead of the nuclear norm due to their equivalence under this setting. Hence, the resulting SDP formulation can be stated as

$$\min_{\mathbf{M}\in\mathbb{R}^{n\times n}} \quad \text{tr}(\mathbf{M}) \qquad \text{s.t. } \mathcal{A}(\mathbf{M}) = b,\ \mathbf{M} \succeq 0, \tag{5}$$

The relation $\delta_{2r} < 1/2$ is a sufficient condition for the recovery of M^* via (5), but not a necessary one (Cai & Zhang, 2013). Recently, Yalcin et al. (2023) showed that a sufficient bound close to $\delta_{2r} \leq 1$ can be achieved when $n \approx 2r$, which proves that the formulation (5) may solve the problem even if the sharp bound $\delta_{2r} \geq 1/2$ for (2) is not satisfied as long as r is a large number (note that the focus of this paper is on the practical scenario of a small rank r).

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 (Wein et al., 2019). The empirical evidence of deep learning practice shows the advantage of using over-parametrized models for both convergence properties during

[1] A point is called spurious if it satisfies first-order and second-order necessary optimality conditions but is not a global minimum.

training (Oymak & Soltanolkotabi, 2020; Zou et al., 2020; Du et al., 2019; Allen-Zhu et al., 2019b) and generalization performance of the trained model (Allen-Zhu et al., 2019a; Neyshabur et al., 2019; Mei & Montanari, 2022; Belkin et al., 2020). Practitioners also design their own hierarchy of machine learning models, to satisfy the scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022; Maloney et al., 2022). 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.

1.2. 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 (Parrilo, 2003; Lasserre, 2001). 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 (2) 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 (2), we could use its dual problem, since the minimum value of (2) 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 (2), where $v := \mathrm{vec}(X)$. Here for simplicity please also 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:

$$\min_{D \in \mathbb{S}^s} \langle F, D \rangle \text{ s.t. } \mathcal{L}(D) = 0, D \succeq 0 \tag{6}$$

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 (6) has some nice properties: it is convex and its optimal value asymptotically reaches that of (2) as κ grows to infinity (under generic conditions), which enables solving the non-convex (2) with an arbitrary accuracy (Lasserre, 2001). However, the problem (6) 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 (2) (without dealing with SDPs) via a lifting technique similar to (6). Currently, there is no guarantee that local minimizers of the BM formulation will translate to the minimizer of the convex problem (6). 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 (Boumal, 2016). 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 (6) 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 (2) via the SOS framework.

1.3. Our Approach

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 (6). 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 (6) could also be casted as an SDP based on the outer product of $v^{\otimes \kappa}$ with itself. The notion of a tensor and its properties (symmetry, rank, etc) will be formally introduced in Section 2 and Appendix A. Instead of solving a non-scalable SDP problem for the optimal D over $\mathbb{S}^s$; 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.

We first focus our attention to the $r = 1$ case, which is easier to conceptualize. The idea is that we replace $\mathbb{R}^n$ with $\mathbb{R}^{n \times \cdots \times n}$, meaning that we replace our decision variable with an l-way tensor $\mathbf{x} \in \mathbb{R}^{n \times \cdots \times n}$. We will show that this new problem can convert spurious solutions of the original problem to strict saddle points in the lifted tensor space, and

we further derive how large l should be in order for this to occur, thereby addressing two main practical deficiencies of (6).

2. Definitions and Notations

The formal definition of a tensor, alongside with the property of symmetry and rank will be elaborated in Appendix A.

Definition 2.1 (Tensor Multiplication)**.** Outer product is an operation carried out on a couple 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 $l+p$, $\mathbf{c} = \mathbf{a} \otimes \mathbf{b}$ with

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

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

$$\underbrace{a \otimes \cdots \otimes a}_{l \text{ times}} \coloneqq a^{\otimes l}$$

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 $q = 1$ and

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

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 $\langle \mathbf{a}, \mathbf{b} \rangle_{q,\dots,s}$.

Definition 2.2 (restricted strong smoothness (RSS))**.** The linear operator $\mathcal{A} : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^m$ satisfies the (L_s, r)-RSS, property if:

$$f(M) - f(N) \leq \langle M - N, \nabla f(N) \rangle + \frac{L_s}{2} \|M - N\|_F^2$$

for all $M, N \in \mathbb{S}^n$ with $\operatorname{rank}(M), \operatorname{rank}(N) \leq r$.

Definition 2.3 (restricted strong convexity (RSC))**.** The linear operator $\mathcal{A} : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^m$ satisfies the (α_s, r)-RSC property if:

$$f(M) - f(N) \geq \langle M - N, \nabla f(N) \rangle + \frac{\alpha_s}{2} \|M - N\|_F^2$$

for all $M, N \in \mathbb{S}^n$ with $\operatorname{rank}(M), \operatorname{rank}(N) \leq r$.

2.1. Notations

In this paper, I_n refers to the identity matrix of size $n \times n$. The notation $M \succeq 0$ means that M is a symmetric and positive semidefinite (PSD) matrix. $\mathbb{S}^n$ denotes the symmetric PSD space of dimension n. $\sigma_i(M)$ denotes the i-th largest singular value of a matrix M, and $\lambda_i(M)$ denotes the i-th largest eigenvalue of M. $\|v\|$ denotes the Euclidean norm of a vector v, while $\|M\|_F$ and $\|M\|_2$ denote the Frobenius norm and induced l_2 norm of a matrix M, respectively. $\langle A, B \rangle$ is defined to be $\operatorname{tr}(A^T B)$ for two matrices A and B of the same size. For a matrix M, $\operatorname{vec}(M)$ is the usual vectorization operation by stacking the columns of the matrix M into a vector. For a vector $v \in \mathbb{R}^{n^2}$, $\operatorname{mat}(v)$ converts v to a square matrix and $\operatorname{mat}_S(v)$ converts v to a symmetric matrix, i.e., $\operatorname{mat}(v) = M$ and $\operatorname{mat}_S(v) = (M + M^T)/2$, where $M \in \mathbb{R}^{n \times n}$ is the unique matrix satisfying $v = \operatorname{vec}(M)$. $[n]$ denotes the integer set $[1, \dots, n]$, and $\circ l$ stands for the shorthand of repeated cartesian product $\times \cdots \times$ for l times. $\mathcal{N}(\mu, \mathbf{\Sigma})$ refers to the multivariate Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$.

3. The Lifted Formulation

To streamline the presentation, we focus on the problem of rank-1 matrix sensing presented in the BM formulation:

$$\min_{x \in \mathbb{R}^n} \quad \|\mathcal{A}(xx^\top - zz^\top)\|_2^2 \tag{7}$$

where $M^* = zz^\top$ is the ground truth rank-1 matrix. The generalization of the ideas to $r > 1$ is straightforward but the mathematical notations will be cumbersome.

The objective is to solve (7) using a lifted or overparametrized 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 (7) aims to find a vector x such that

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

Therefore, it is also desirable to achieve

$$\{\mathcal{A}(xx^\top)\}^{\otimes l} = b^{\otimes l} \in \mathbb{R}^{m \circ l}$$

According to (Petersen et al., 2008), for arbitrary 4 vectors a, b, c, d of the same dimension it holds that

$$\langle a \otimes b, c \otimes d \rangle = \langle a, c \rangle \langle b, d \rangle \tag{8}$$

With the repeated application of the above identity, we have

$$\{\mathcal{A}(xx^\top)\}^{\otimes l} = \langle \mathbf{A}, x^{\otimes l} \otimes x^{\otimes l} \rangle \tag{9}$$

where the tensor $\mathbf{A} \in \mathbb{R}^{(m \circ l) \times (n^2 \circ l)}$ is defined as:

$$A_{m_1 \dots m_l} = \prod_{k=1}^{l} \otimes \operatorname{vec}(A_{m_k}). \tag{10}$$

Therefore, one can write the lifted objective similarly to (7) as:

$$\min_{\mathbf{w}\in\mathbb{R}^{n^o l}} \quad \|\langle \mathbf{A}, \mathbf{w}\otimes\mathbf{w} - z^{\otimes l}\otimes z^{\otimes l}\rangle_{m^l+1,\ldots,m^l+n^{2l}}\|_F^2 \quad (11)$$

For notational convenience, define $f(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}$ and $h(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}$ as:

$$f(M) := \|\mathcal{A}(M - zz^\top)\|_2^2, \qquad h(x) = f(xx^\top),$$

and $\nabla f(\cdot) = \nabla_M f(\cdot)$ and $\nabla h(\cdot) = \nabla_x h(\cdot)$.

Similarly, define $f^l(\cdot) : \mathbb{R}^{n^o 2l} \mapsto \mathbb{R}$ and $h^l(\cdot) : \mathbb{R}^{n^o l} \mapsto \mathbb{R}$ as:

$$\begin{aligned} &f^l(\mathbf{M}) := \|\langle \mathbf{A}, \mathbf{M} - z^{\otimes l}\otimes z^{\otimes l}\rangle_{m^l+1,\ldots,m^l+n^{2l}}\|_F^2, \\ &h^l(\mathbf{w}) = f^l(\mathbf{w}\otimes\mathbf{w}), \end{aligned} \quad (12)$$

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

4. 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

$$\mathcal{A}_\epsilon(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij}, & \text{if } (i,j)\in\Omega \\ \epsilon\mathbf{M}_{ij}, & \text{otherwise} \end{cases}, \quad (13)$$

where Ω is a measurement set such that

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

Yalcin et al. (2023) 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 $\delta_{2r} = (1-\epsilon)/(1+\epsilon)$ for some sufficiently small ϵ.

To study whether our lifted framework can reshape the optimization landscape of the problem, we analyze the spurious local minima of the unlifted problem (7). Given any spurious local minimum $\hat{x}$, it is essential to understand whether its lifted counterpart $\hat{x}^{\otimes l}$ behaves differently in (11), 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 (7) for the measurement matrices given in (13), and then find the smallest eigenvalue of the Hessian of (11) at the lifted counterpart of each spurious solution. We summarize the findings in Table 1 for $\epsilon = 0.3$. Note that due to the structure of (13), 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.

It can be observed that, for a given spurious local minimizer $\hat{x}$ of (7), 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. To further study how much parametrization is needed and to show that this is not unique to any particular problem form, we provide theoretical results next.

5. Optimization Landscape of the Lifted Problem

We analyze the optimization landscape of (11) around $\hat{x}^{\otimes l}$, where $\hat{x}$ is a spurious spurious solution of (7). The proofs to all of the results below can be found in Appendix B.

5.1. FOP and SOP conditions

Lemma 5.1. *The vector $\hat{x}$ is a SOP of (7) if and only if*

$$\nabla f(\hat{x}\hat{x}^\top)\hat{x} = 0, \quad (14a)$$

$$\begin{aligned} &2\langle \nabla f(\hat{x}\hat{x}^\top), uu^\top\rangle + \\ &[\nabla^2 f(\hat{x}\hat{x}^\top)](\hat{x}u^\top + u\hat{x}^\top, \hat{x}u^\top + u\hat{x}^\top) \geq 0 \end{aligned} \quad (14b)$$

$\forall u \in \mathbb{R}^n$, with (14a) being the necessary and sufficient condition for $\hat{x}$ to be a first-order point (FOP), which is a stationary point of the objective function.

Please refer to (Ha et al., 2020) or (Zhang et al., 2021) for a proof of this lemma. Lemma 5.1 has a counterpart in the lifted space since (7) and (11) are highly similar. This will be formalized below.

Lemma 5.2. *The tensor $\hat{\mathbf{w}}$ is a SOP of (11) if and only if*

$$\langle \nabla f^l(\hat{\mathbf{w}}\otimes\hat{\mathbf{w}}), \hat{\mathbf{w}}\rangle_{n^l+1,\ldots,n^{2l}} = 0, \quad (15a)$$

$$\begin{aligned} &2\langle \nabla f^l(\hat{\mathbf{w}}\otimes\hat{\mathbf{w}}), \Delta\otimes\Delta\rangle + [\nabla^2 f^l(\hat{\mathbf{w}}\otimes\hat{\mathbf{w}})] \\ &(\hat{\mathbf{w}}\otimes\Delta + \Delta\otimes\hat{\mathbf{w}}, \hat{\mathbf{w}}\otimes\Delta + \Delta\otimes\hat{\mathbf{w}}) \geq 0 \quad \forall \Delta\in\mathbb{R}^{n^o l} \end{aligned} \quad (15b)$$

with (15a) being a necessary and sufficient condition for $\hat{x}$ to be a FOP.

Table 1. The smallest eigenvalue of the Hessian of lifted SOPs of (7)

n	l	$\|\nabla h^l(z^{\otimes l})\|_F$	$\|\nabla h^l(\hat{x}^{\otimes l})\|_F$	$\lambda_{\min}(\nabla^2 h^l(z^{\otimes l}))$	$\lambda_{\min}(\nabla^2 h^l(\hat{x}^{\otimes l}))$
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

5.2. Optimality condition of lifted problem, symmetric rank-1 constraint

Theorem 5.3. *For* (11)*, the equality* $\nabla h^l(\hat{\mathbf{w}}) = 0$ *holds if*

$$\hat{\mathbf{w}} = \hat{x}^{\otimes l}$$

where $\hat{x} \in \mathbb{R}^n$ *is an FOP of* (7)*.*

Theorem 5.3 theoretically confirms the phenomenon that we observed in the numerical example above, by asserting that all FOPs of the unlifted problem (7) are still FOPs in the lifted domain by transforming this point using tensor outer product, or by overparametrization. This means that some critical geometric structures of (7) are still maintained in (10), establishing strong connections between these two representations of the same problem.

After establishing the above negative result about a FOC having remained a stationary point after lifting, we turn to studying the differences between (7) and (11), since Table 1 suggests that the negative curvature of the Hessian at each spurious local minimizer will disappear after enough over-parametrization. Thus, the central problem under study is that whether for an instance of (7) with spurious local minima, these undesirable points will continue to be spurious solutions in the lifted space. If so, there is no apparent benefit to performing the lifting operation. Conversely, if the stationary points obtain a negative curvature, one can select from a large set of low-complexity algorithms to efficiently escape from strict saddles, and therefore eliminate local minima from the problem formulation.

The following theorem demonstrates that lifting enables the elimination of spurious solutions, and we can further derive a bound on the order l needed to achieve the elimination of spurious solutions given the RSS and RSC constants of the problem.

Theorem 5.4. *Consider a SOP* $\hat{x} \neq \pm z \in \mathbb{R}^n$ *of* (7)*, and assume that* (7) *satisfies the RSC and RSS conditions. Then* $\hat{\mathbf{w}} = \hat{x}^{\otimes l}$ *is a strict saddle of* (11) *with a rank-1 symmetric escape direction if* $\hat{x}$ *satisfies the inequality*

$$\|M^* - \hat{x}\hat{x}^\top\|_F^2 \geq \frac{L_s}{\alpha_s}\|\hat{x}\|_2^2 \operatorname{tr}(M^*) \tag{16}$$

and l *is odd and is large enough so that*

$$l > \frac{1}{1 - \log_2(2\beta)} \tag{17}$$

where β *is defined as*

$$\beta := \frac{L_s \operatorname{tr}(M^*)\|\hat{x}\|_2^2}{\alpha_s \|M^* - \hat{x}\hat{x}^\top\|_F^2}.$$

Here, L_s, α_s *are the respective RSS and RSC constants of* (7)*.*

Theorem 5.4 is powerful since it proves that by lifting spurious solutions to higher-order tensor spaces, we can convert them to saddle points, which attests to the power of over-parametrization. More importantly, regardless of how large L_s/α_s is, it is always possible to find an order l large enough to convert $\hat{x}^{\otimes l}$s to saddle points, which is a major improvement over the existing results such as Theorem 1.2. Equation (16) implies that in order for $\hat{x}^{\otimes l}$ to become a saddle point in the lifted formulation, it is enough to have either a small $\hat{x}$ or a large $\|M^* - \hat{x}\hat{x}^\top\|_F^2$. As a result, no spurious solution $\hat{x}$ close to the origin will remain spurious after lifting. This is a significant result because if we initialize a saddle-escaping algorithm near the origin, it will not be trapped inside a spurious solution during the early iterations even if the problem is highly non-convex. One fact to note is that (16) is not a necessary condition, but a sufficient one, and therefore it is possible that the statements of Theorem 5.4 can be applied to a wider range of $\hat{x}$.

Another major advantage of Theorem 5.4 is that it quantifiably studies how many levels of parametrization are need in order to make an existing spurious local minimizer $\hat{x}$ a saddle point in the lifted space. Therefore, instead of offering a statement asserting that over-parametrization will work at some large enough l (as done in the SOS setting), it explains how large this l needs to be in terms of its geometric

regularities, captured by the RSC and RSS parameters, and also the distance $\|M^* - \hat{x}\hat{x}^\top\|_F^2$.

Theorem 5.4 also implies that over-parametrization works particularly well for those spurious solutions $\hat{x}$ far away from the ground truth, in the sense that the further away $\hat{x}$ is from the ground truth, the smaller l needs to be in order for $\hat{x}^{\otimes l}$ to become a saddle point, as suggested by (17). This fact is in line with the existing literature saying that the optimization landscape near M^* is benign, in the sense that there exists no spurious local solution in a region around M^*. A well-known incarnation of the aforementioned statement is given below.

Theorem 5.5 (Theorem 3 (Bi & Lavaei, 2020))**.** *If $\hat{x}$ is a SOP of (7) and*

$$\|\hat{x}\hat{x}^\top - M^*\|_F \leq \frac{4L_s\alpha_s}{(L_s+\alpha_s)^2}\|z\|_2^2, \tag{18}$$

then

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

This means that any spurious solution must violate the inequality (18). This allows us to simplify the results of Theorem 5.4, which can be stated in the following form

Theorem 5.6. *Assume that $\hat{x} \neq \pm z \in \mathbb{R}^n$ is a spurious solution of (7), and that (7) satisfies the RSC and RSS assumptions with α_s and L_s constants respectively. The point $\hat{x}^{\otimes l}$ will become a saddle point of (11) for an odd l satisfying (17) if*

$$\|M^*\|_F \leq \frac{2\sqrt{2}\alpha_s^{5/2}}{(L_s+\alpha_s)^2\sqrt{L_s}} \tag{19}$$

This theorem is proved by setting the RHS of (18) to be smaller than that of (16). Another technical lemma is introduced to bridge the two terms, so that it does not depend on specific $\hat{x}$ anymore.

The above results all aim to convert spurious solutions to saddle points via lifting. Although this property is highly desirable for spurious local solutions, it is essential to make sure that it will not hold for the ground truth solution since the correct solution should remain a SOP in the lifted space in order for the lifting technique to be useful.

In our previous numerical experiment, we empirically showed that the smallest eigenvalue of the Hessian at $z^{\otimes l}$ remains positive, meaning that it is still a SOP in the lifted formulation (11). In the following theorem we formally establish this observation.

Theorem 5.7. *Assume that $z \in \mathbb{R}^n$ is the ground truth solution of (7). Then $z^{\otimes l}$ remains a SOP of (11) regardless of the parametrization level l, and without the need for (7) to satisfy RSC or RSS conditions.*

6. Numerical Experiments

In this section, we numerically demonstrate that the theoretical results of this paper can be translated to real advantages when using the lifted framework (11)[2].

For the sake of convenience, we revisit the matrix sensing problem (7) with $n = 3$ and the special operator (13). We choose $\epsilon = 0.3$ 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 (13), it is easy to verify that there are theoretically 4 SOPs in total, and they converge to the following 4 points as ϵ becomes sufficiently small, which are:

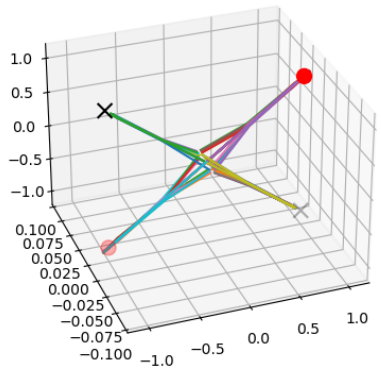
$$\hat{x}_1 \approx \begin{bmatrix}1\\0\\1\end{bmatrix}, \hat{x}_2 \approx \begin{bmatrix}1\\0\\-1\end{bmatrix}, \hat{x}_3 \approx \begin{bmatrix}-1\\0\\1\end{bmatrix}, \hat{x}_4 \approx \begin{bmatrix}-1\\0\\-1\end{bmatrix}.$$

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.

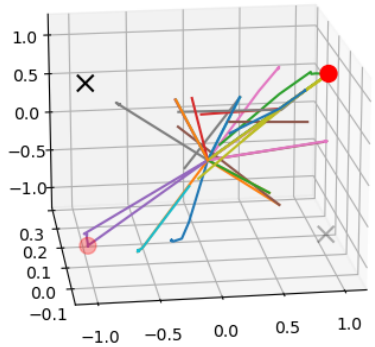
To empirically verify that $\epsilon = 0.3$ is indeed small enough, we simply start from random Gaussian initialization, and apply optimization algorithms to check to what point(s) the algorithm will eventually converge to. We use the standard ADAM optimizer (Kingma & Ba, 2014) with the hyperparameter lr $= 0.02$, and the 3D convergence trajectories are plotted in Figure 1(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 $\epsilon = 0.3$ 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 (7) with (13) 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 (13).

Thus, at a high level, it is necessary to show that by using the lifted framework (11), 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 5.4. To this

[2]The code used to produce results in this section can be found at: https://github.com/anonpapersbm/liftedmatrixsensing



(a) Convergence trajectories of unlifted formulation.



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

Figure 1. The convergence trajectories of (13), with $n = 3, \epsilon = 0.3$. Random gaussian initialization with $\sigma = 0.01, \mu = 0$. 40 Trials in total.

end, we plot the optimization trajectory of (11) with $l = 3$ and (13) in Figure 1(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 1(a) meaningful. The reason that we chose $l = 3$ instead of $l = 2$ is because Theorem 5.4 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 1(b) seems more complicated than Figure 1(a), as an extra layer of approximation is applied. Nevertheless, the message of Figure 1(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 1(a). To further demonstrate the power of the over-parametrized framework (11), we summarize the success rate of unlifted framework (7) and the lifted framework (11) in the table below. Here, we count a trial to be a "success" if the final iteration x_{terminal} satisfies

$$\|x_{\text{terminal}}^{\otimes l} - z^{\otimes l}\|_F \leq 0.05$$

From Table 2, we can see that n increases, the success rate of the lifted framework goes up, especially in contrast to the fact that higher n means lower success rate for the unlifted formulation due to it having $\mathcal{O}(2^{\lceil n/2 \rceil} - 2)$ spurious local solution. This empirically demonstrates that the lifted formulation is especially valuable in problems with higher dimensions.

Table 2. Success rate of the lifted and unlifted frameworks applied to (13)

	Unlifted $l = 1$	Lifted $l = 3$
$n = 3$	0.575	0.62
$n = 4$	0.575	0.68
$n = 5$	0.475	0.75

7. Conclusion

This paper 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. Future research directions include the sparsification of the lifting method to eliminate unnecessary monomials and reduce the complexity, as well as studying whether lifting will create new stationary points and where they are located relative to the ground truth solution.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019a.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019b.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

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

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

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Cai, T. T. and Zhang, A. Sharp rip bound for sparse signal and low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 35(1):74–93, 2013.

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

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

Comon, P., Golub, G., Lim, L.-H., and Mourrain, B. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1254–1279, 2008.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

Fattahi, S. and Sojoudi, S. 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.

Ha, W., Liu, H., and Barber, R. F. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4): 2927–2955, 2020.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kolda, T. G. Numerical optimization for symmetric tensor decomposition. *Mathematical Programming*, 151(1):225–248, 2015.

Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37, 2009.

Lasserre, J. B. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

Levin, E., Kileel, J., and Boumal, N. The effect of smooth parametrizations on nonconvex optimization landscapes. *arXiv preprint arXiv:2207.03512*, 2022.

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

Ma, Z., Bi, Y., Lavaei, J., and Sojoudi, S. 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, pp. 7672–7681, 2022.

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

Maloney, A., Roberts, D. A., and Sully, J. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Mei, S. and Montanari, A. 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.

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

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.

Oymak, S. and Soltanolkotabi, M. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Parrilo, P. A. Semidefinite programming relaxations for semialgebraic problems. *Mathematical programming*, 96 (2):293–320, 2003.

Petersen, K. B., Pedersen, M. S., et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

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

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

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

Wein, A. S., El Alaoui, A., and Moore, C. The kikuchi hierarchy and tensor pca. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1446–1468. IEEE, 2019.

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

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

Zhang, R. Y. Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime. *arXiv preprint arXiv:2104.10790*, 2021.

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

Zhang, R. Y., Lavaei, J., and Baldick, R. Spurious local minima in power system state estimation. *IEEE transactions on control of network systems*, 6(3):1086–1096, 2019a.

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

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

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109(3):467–492, 2020.

A. Definition

Definition A.1 (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 (Comon et al., 2008). In particular, we define an l-way array as such:

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

Note that in this paper tensors and arrays can be regarded as synonymous since there exists an isomorphism between them. Moreover, if $n_1 = \cdots = n_l$, 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 $n \circ l := n \times \cdots \times n$. In this work, tensors are denoted with bold variables, and other fonts are reserved for matrices, vectors, and scalars unless specified otherwise.

Definition A.2 (Symmetric Tensor)**.** Similar to the definition of symmetric matrices, for an order-l tensor $\mathbf{a}$ with the same dimensions (i.e., $n_1 = \cdots = n_l$), 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 A.3 (Rank of Tensors)**.** The rank of a cubic tensor $\mathbf{a} \in \mathbb{R}^{n \circ l}$ is defined as:

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

where $u_i, \dots, w_i \in \mathbb{R}^n \ \forall i$. Furthermore, according to (Kolda, 2015), 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 nonnegative λ_is, which is very similar to the rank of symmetric matrices indeed. For notational convenience, we denote rank-r symmetric tensors as $\mathrm{S}^l(\mathbb{R}^n)_r$.

B. Proofs

Proof of Theorem 5.3. (14a) implies that

$$\sum_{a,i,j,s}^{[m]\times[n]\times[n]\times[n]} (A_a)_{sk} (A_a)_{ij} \hat{x}_i \hat{x}_j \hat{x}_s = \sum_{a,i,j,s}^{[m]\times[n]\times[n]\times[n]} (A_a)_{sk} (A_a)_{ij} z_i z_j \hat{x}_s \qquad \forall k \tag{20}$$

Then we focus on (15a) with

$$\nabla f^l(\hat{\mathbf{w}} \otimes \hat{\mathbf{w}}) = \langle \mathbf{A}^\top \mathbf{A}, \hat{\mathbf{w}} \otimes \hat{\mathbf{w}} - z^{\otimes l} \otimes z^{\otimes l} \rangle_{n^{2l}+1, \dots, 2n^{2l}}$$

where $\mathbf{A}^\top \mathbf{A} := \langle \mathbf{A}, \mathbf{A} \rangle_{m^l+1, \dots, m^l+n^{2l}}$. Thus, the LHS of (15a) is:

$$\begin{aligned} &\sum_{\{a_\alpha, i_\alpha, j_\alpha, s_\alpha\}_{\alpha=1}^{l}}^{([m]\circ l)\times([n]\circ l)\times([n]\circ l)\times([n]\circ l)} \left(\prod_{\alpha=0}^{l} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} \hat{x}_{i_\alpha} \hat{x}_{j_\alpha} \hat{x}_{s_\alpha} \right) - \left(\prod_{\alpha=0}^{l} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} z_{i_\alpha} z_{j_\alpha} \hat{x}_{s_\alpha} \right) \\ = &\prod_{\alpha=0}^{l} \left(\sum_{a_\alpha, i_\alpha, j_\alpha, s_\alpha}^{[m]\times[n]\times[n]\times[n]} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} \hat{x}_{i_\alpha} \hat{x}_{j_\alpha} \hat{x}_{s_\alpha} \right) - \prod_{\alpha=0}^{l} \left(\sum_{a_\alpha, i_\alpha, j_\alpha, s_\alpha}^{[m]\times[n]\times[n]\times[n]} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} z_{i_\alpha} z_{j_\alpha} \hat{x}_{s_\alpha} \right) \end{aligned} \tag{21}$$

Given (20) and (10), we know that:

$$\sum_{a_\alpha, i_\alpha, j_\alpha, s_\alpha}^{[m]\times[n]\times[n]\times[n]} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} \hat{x}_{i_\alpha} \hat{x}_{j_\alpha} \hat{x}_{s_\alpha} = \sum_{a_\alpha, i_\alpha, j_\alpha, s_\alpha}^{[m]\times[n]\times[n]\times[n]} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} z_{i_\alpha} z_{j_\alpha} \hat{x}_{s_\alpha} \quad \forall k_\alpha$$

Therefore, substituting the above equality into (21) yields that LHS of (15a) is 0.

□

Before proceeding to the proof of Theorem 5.4, we first recall a useful technical Lemma from (Ma & Sojoudi, 2023):

Lemma B.1. *For any SOP $\hat{x}$ of (7), define G as $G := -\lambda_{min}(\nabla f(\hat{x}\hat{x}^\top))$, and L_s be the RSS constant. Then it holds that:*

$$G \leq \|\hat{x}\|_2^2 L_s$$

Proof of Theorem 5.4. (14b) implies that:

$$\begin{aligned}\text{LHS} = 2 \sum_{a,i,j,s,k}^{[m]\times[n]\circ 4} (A_a)_{sk}(A_a)_{ij}(\hat{x}_i\hat{x}_j - z_i z_j)u_k u_s + \\ \sum_{a,i,j,s,k}^{[m]\times[n]\circ 4} (A_a)_{sk}(A_a)_{ij}(\hat{x}_i u_j + u_i\hat{x}_j)(\hat{x}_s u_k + u_s\hat{x}_k)\end{aligned} \tag{22}$$

If the A_a matrices are symmetric, which can be achieved by redefining A_a as $(A_a^T + A_a)/2$ without changing the measurement values, the above equation can be simplified as:

$$\underbrace{2 \sum_{a,i,j,s,k}^{[m]\times[n]\circ 4} (A_a)_{sk}(A_a)_{ij}(\hat{x}_i\hat{x}_j - z_i z_j)u_k u_s}_{C_1} + \underbrace{4 \sum_{a,i,j,s,k}^{[m]\times[n]\circ 4} (A_a)_{sk}(A_a)_{ij}\hat{x}_i\hat{x}_k u_j u_s}_{C_2} \tag{23}$$

According to (Zhang et al., 2021), $\nabla f(M)$ can be assumed to be symmetric without loss of generality. Hence, one can select $u \in \mathbb{R}^n$ such that $u^\top \nabla f(\hat{x}\hat{x}^\top)u = \lambda_{\min}(\nabla f(\hat{x}\hat{x}^\top))$ and $\lambda_{\min}(\nabla f(\hat{x}\hat{x}^\top)) \leq 0$ under the RSC assumption with $\alpha_s \geq 0$. The reason that this holds is because first we know that

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

Since $\langle \nabla f(\hat{x}\hat{x}^\top), \hat{x}\hat{x}^\top \rangle = 0$ according to (14a) and $f(\hat{x}\hat{x}^\top) - f(M^*) \geq 0$, we know that

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

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

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

which implies that

$$\lambda_{\min}(\nabla f(\hat{x}\hat{x}^\top)) \leq -\alpha_s \frac{\|\hat{x}\hat{x}^\top - M^*\|_F^2}{\operatorname{tr}(M^*)} \leq 0 \tag{24}$$

With this piece of knowledge in mind, we define $G := -\lambda_{\min}(\nabla f(\hat{x}\hat{x}^\top)) \geq 0$. Thus,

$$C_1 = -G.$$

Moreoever, the RSS condition implies that:

$$\begin{aligned}4C_2 &= [\nabla^2 f(\hat{x}\hat{x}^\top)](\hat{x}u^\top + u\hat{x}^\top, \hat{x}u^\top + u\hat{x}^\top) \leq L_s \|\hat{x}u^\top + u\hat{x}^\top\|_F^2 \\ &= L_s \operatorname{tr}((\hat{x}u^\top + u\hat{x}^\top)^\top \hat{x}u^\top + u\hat{x}^\top) = 2L_s\|\hat{x}\|_2^2\end{aligned}$$

since $u^\top \hat{x} = 0$ according to the first-order condition (14a). Therefore,

$$C_2 \leq \frac{1}{2} L_s \|\hat{x}\|_2^2$$

Now, we take a look at the left-hand side (LHS) of (15b); here we choose $\Delta = u^{\otimes l}$ for the same $u \in \mathbb{R}^n$ chosen above:

$$\left[\underbrace{2 \sum_{\{a_\alpha, i_\alpha, j_\alpha, s_\alpha, k_\alpha\}_{\alpha=1}^{l}}^{([m]\circ l)\times[([n]\circ l)\circ 4]} \left(\prod_{\alpha=0}^{l} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} \hat{x}_{i_\alpha} \hat{x}_{j_\alpha} u_{s_\alpha} u_{k_\alpha} \right) - \left(\prod_{\alpha=0}^{l} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} z_{i_\alpha} z_{j_\alpha} u_{s_\alpha} u_{k_\alpha} \right)}_{C_3} \right] + \underbrace{4 \sum_{\{a_\alpha, i_\alpha, j_\alpha, s_\alpha, k_\alpha\}_{\alpha=1}^{l}}^{([m]\circ l)\times[([n]\circ l)\circ 4]} \prod_{\alpha=0}^{l} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} \hat{x}_{i_\alpha} \hat{x}_{k_\alpha} u_{j_\alpha} u_{s_\alpha}}_{C_4} \tag{25}$$

Now,

$$\begin{aligned} C_3 &= 2 \prod_{\alpha=1}^{l} \left(\sum_{a,i,j,s,k}^{[m]\times[n]\circ 4} A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} \hat{x}_{i_\alpha} \hat{x}_{j_\alpha} u_{s_\alpha} u_{k_\alpha} \right) - 2 \prod_{\alpha=1}^{l} \left(\sum_{a,i,j,s,k}^{[m]\times[n]\circ 4} (A_{a_\alpha s_\alpha k_\alpha} A_{a_\alpha i_\alpha j_\alpha} z_{i_\alpha} z_{j_\alpha} u_{s_\alpha} u_{k_\alpha} \right) \\ &= 2 \left(\sum_{a,i,j,s,k}^{[m]\times[n]\circ 4} (A_a)_{sk} (A_a)_{ij} \hat{x}_i \hat{x}_j u_k u_s \right)^l - 2 \left(\sum_{a,i,j,s,k}^{[m]\times[n]\circ 4} (A_a)_{sk} (A_a)_{ij} z_i z_j u_k u_s \right)^l \\ &\leq 2 \left(\sum_{a,i,j,s,k}^{[m]\times[n]\circ 4} (A_a)_{sk} (A_a)_{ij} (\hat{x}_i \hat{x}_j - z_i z_j) u_k u_s \right)^l = 2C_1^l = -2G^l \end{aligned} \tag{26}$$

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. Next,

$$C_4 = \left(\sum_{a,i,j,s}^{[m]\times[n]\circ 4} (A_a)_{sk} (A_a)_{ij} \hat{x}_i \hat{x}_k u_j u_s \right)^l = C_2^l \leq \frac{1}{2^l} L_s^l \|\hat{x}\|_2^{2l} \tag{27}$$

As a result,

$$\text{LHS of (15b)} \leq \underbrace{-2G^l}_{\text{Part 1}} + \underbrace{\frac{2}{2^{l-1}} L_s^l \|\hat{x}\|_2^{2l}}_{\text{Part 2}}$$

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

$$G^l > (1/2^{l-1}) L_s^l \|\hat{x}\|_2^{2l} \tag{28}$$

to be able to make the LHS of (15b) negative.

To derive a sufficient condition for (28), we first need a lowerbound on G, and to do so, we start with the (sparse) RSC assumption:

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

Then, since $f(M^*) \leq f(\hat{x}\hat{x}^\top)$ (as M^* is the global optimum), we have:

$$0 \geq \langle \nabla f(\hat{x}\hat{x}^\top), M^* - \hat{x}\hat{x}^\top \rangle + \frac{\alpha_s}{2} \|M^* - \hat{x}\hat{x}^\top\|_F^2 = \langle \nabla f(\hat{x}\hat{x}^\top), M^* \rangle + \frac{\alpha_s}{2} \|M^* - \hat{x}\hat{x}^\top\|_F^2 \tag{29}$$

where the equality follows from the FOP condition (14a) for $\hat{x}$. Also, since M^* is assumed to be positive semidefinite, we have

$$\langle \nabla f(\hat{x}\hat{x}^\top), M^* \rangle \geq \lambda_{\min}(\nabla f(\hat{x}\hat{x}^\top)) \operatorname{tr}(M^*). \tag{30}$$

Combining (29) and (30), we have:

$$-G = \lambda_{\min}(\nabla f(\hat{x}\hat{x}^\top)) \leq -\frac{\alpha_s}{2\,\mathrm{tr}(M^*)}\|M^* - \hat{x}\hat{x}^\top\|_F^2,$$

meaning that

$$G \geq \frac{\alpha_s}{2\,\mathrm{tr}(M^*)}\|M^* - \hat{x}\hat{x}^\top\|_F^2 \tag{31}$$

Therefore, if

$$\left(\frac{\alpha_s}{2\,\mathrm{tr}(M^*)}\|M^* - \hat{x}\hat{x}^\top\|_F^2\right)^l > (1/2^{l-1})L_s^l\|\hat{x}\|_2^{2l},$$

we can conclude that (28) holds, which implies that the LHS of (15b) is negative, directly proving that $\hat{x}^{\otimes l}$ is not a SOP anymore. Elementary manipulations of the above equation give that a sufficient condition is

$$\|M^* - \hat{x}\hat{x}^\top\|_F^2 > 2^{1/l}\frac{L_s}{\alpha_s}\|\hat{x}\|_2^2\,\mathrm{tr}(M^*) \tag{32}$$

We now consider (16), which means that

$$\|\hat{x}\|^2 \leq \frac{\alpha_s}{L_s\,\mathrm{tr}(M^*)}\|M^* - \hat{x}\hat{x}^\top\|_F^2 \tag{33}$$

Subsequently, define a constant γ such that:

$$L_s\|\hat{x}\|_2^2 = \gamma(\frac{\alpha_s}{2\,\mathrm{tr}(M^*)}\|M^* - \hat{x}\hat{x}^\top\|_F^2)$$

Then according to Lemma B.1 and (31), we can conclude that $\gamma \geq 1$. Moreover, (33) also means that $\gamma < 2$. So with this new definition, the sufficient condition (32) becomes

$$1 > \frac{\gamma}{2^{(l-1)/l}} \tag{34}$$

Since we already know that $1 \leq \gamma < 2$, there always exists a large enough l such that (34) holds, which in turn implies that LHS of (15b) is negative, proving that $\hat{x}^{\otimes l}$ is a saddle point with the escape direction $u^{\otimes l}$, proving the claim.

Next, we aim to study how large l needs to be in order for (34) to hold. Now, by utilizing Lemma B.2 again, we know that

$$\gamma = \frac{2L_s\,\mathrm{tr}(M^*)\|\hat{x}\|_2^2}{\alpha_s\|M^* - \hat{x}\hat{x}^\top\|_F^2} := 2\beta$$

and we know $\beta \leq 1$ due to assumption (16). So for (34) to hold true, we need

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

□

Proof of Theorem 5.6. First, consider the following technical lemma, which is proved below this proof,

Lemma B.2. *Given a FOP $\hat{x}$ of (7), it holds that*

$$\|\hat{x}\|^2 < \sqrt{\frac{2L_s}{\alpha_s}}\|M^*\|_F \tag{35}$$

Via the above lemma, we know that a sufficient condition to (16) is

$$\|M^* - \hat{x}\hat{x}^\top\|_F^2 \geq \sqrt{\frac{2L_s^3}{\alpha_s^3}}\|M^*\|_F\,\mathrm{tr}(M^*)$$

Making the RHS of the above inequality to be smaller than RHS of (18) proves the theorem, especially by acknowledging that M^* is rank-1. □

Proof of Lemma B.2. Lemma 6 of (Zhang et al., 2021) states that given an arbitrary constant λ and vector $u \in \mathbb{R}^n$,

$$\|u\|_2^4 \geq \max\{\frac{2L_s}{\alpha_s}\|M^*\|_F^2, (\frac{2\lambda\sqrt{r}}{\alpha_s})^{4/3}\} \implies \|\nabla h(u)\|_F \geq \lambda$$

A simple negation to both sides gives

$$\|\nabla h(u)\|_F < \lambda \implies \|u\|_2^4 < \max\{\frac{2L_s}{\alpha_s}\|M^*\|_F^2, (\frac{2\lambda\sqrt{r}}{\alpha_s})^{4/3}\}$$

If we set $u = \hat{x}$, then LHS of the above relationship is automatically satisfied for arbitrarily small λ since $\|\nabla h(\hat{x})\|_F = 0$, and thus we conclude that

$$\|u\|_2^4 < \frac{2L_s}{\alpha_s}\|M^*\|_F^2$$

since $(\frac{2\lambda\sqrt{r}}{\alpha_s})^{4/3}$ can be made arbitrarily small. □

Proof of Theorem 5.7. Again utilizing the assumption that A_s matrices are symmetric(A_s can be converted to be symmetric without altering the observation b), we arrive at

$$\text{LHS of (14b)} = 2\sum_{a,i,j,s}^{[m]\times[n]\circ 4}(A_a)_{sk}(A_a)_{ij}(\hat{x}_i\hat{x}_j - z_iz_j)u_ku_s + 4\sum_{a,i,j,s}^{[m]\times[n]\circ 4}(A_a)_{sk}(A_a)_{ij}\hat{x}_i\hat{x}_ku_ju_s \geq 0 \tag{36}$$

for any SOP $\hat{x}$. If we substitute z into the above equation, we obtain that for any $u \in \mathbb{R}^n$

$$\sum_{a,i,j,s}^{[m]\times[n]\circ 4}(A_a)_{sk}(A_a)_{ij}z_iz_ku_ju_s \geq 0 \tag{37}$$

Then given any $\Delta \in \mathbb{R}^{n\otimes l}$, we CP decompose (CANDECOMP, a standard tensor decomposition scheme) it as:

$$\Delta = \sum_{p=1}^{R}\delta^{p,1}\otimes\cdots\otimes\delta^{p,l}$$

where R is the rank of Δ, a finite number. Next, we consider (15b) evaluated at $z^{\otimes l}$, and we have that LHS of (15b) equals:

$$\begin{aligned}
&\sum_{p=1}^{R}\left[2\sum_{\{a_\alpha,i_\alpha,j_\alpha,s_\alpha,k_\alpha\}_{\alpha=1}^{l}}^{([m]\circ l)\times[([n]\circ l)\circ 4]}\left(\prod_{\alpha=0}^{l}A_{a_\alpha s_\alpha k_\alpha}A_{a_\alpha i_\alpha j_\alpha}z_{i_\alpha}z_{j_\alpha}\delta_{s_\alpha}^{p,\alpha}\delta_{k_\alpha}^{p,\alpha}\right) - \left(\prod_{\alpha=0}^{l}A_{a_\alpha s_\alpha k_\alpha}A_{a_\alpha i_\alpha j_\alpha}z_{i_\alpha}z_{j_\alpha}\delta_{s_\alpha}^{p,\alpha}\delta_{k_\alpha}^{p,\alpha}\right)\right] + \\
&4\sum_{\{a_\alpha,i_\alpha,j_\alpha,s_\alpha,k_\alpha\}_{\alpha=1}^{l}}^{([m]\circ l)\times[([n]\circ l)\circ 4]}\prod_{\alpha=0}^{l}A_{a_\alpha s_\alpha k_\alpha}A_{a_\alpha i_\alpha j_\alpha}z_{i_\alpha}z_{k_\alpha}\delta_{j_\alpha}^{p,\alpha}\delta_{s_\alpha}^{p,\alpha} \\
&= \sum_{p=1}^{R}4\sum_{\{a_\alpha,i_\alpha,j_\alpha,s_\alpha,k_\alpha\}_{\alpha=1}^{l}}^{([m]\circ l)\times[([n]\circ l)\circ 4]}\prod_{\alpha=0}^{l}A_{a_\alpha s_\alpha k_\alpha}A_{a_\alpha i_\alpha j_\alpha}z_{i_\alpha}z_{k_\alpha}\delta_{j_\alpha}^{p,\alpha}\delta_{s_\alpha}^{p,\alpha} \\
&= 4\sum_{p=1}^{R}\prod_{\alpha=0}^{l}\left(\sum_{a,i,j,s}^{[m]\times[n]\circ 4}A_{a_\alpha s_\alpha k_\alpha}A_{a_\alpha i_\alpha j_\alpha}z_{i_\alpha}z_{k_\alpha}\delta_{j_\alpha}^{p,\alpha}\delta_{s_\alpha}^{p,\alpha}\right) \geq 0
\end{aligned} \tag{38}$$

where the last inequality follows from (37).

□