Algorithmic Regularization in Tensor Optimization: Towards a Lifted Approach in Matrix Sensing

Ziye Ma
Department of EECS
UC Berkeley
ziyema@berkeley.edu

Javad Lavaei
Department of IEOR
UC Berkeley
lavaei@berkeley.edu

Somayeh Sojoudi
Department of EECS, ME
UC Berkeley
sojoudi@berkeley.edu

Abstract

Gradient descent (GD) is crucial for generalization in machine learning models, as it induces implicit regularization, promoting compact representations. In this work, we examine the role of GD in inducing implicit regularization for tensor optimization, particularly within the context of the lifted matrix sensing framework. This framework has been recently proposed to address the non-convex matrix sensing problem by transforming spurious solutions into strict saddles when optimizing over symmetric, rank-1 tensors. We show that, with sufficiently small initialization scale, GD applied to this lifted problem results in approximate rank-1 tensors and critical points with escape directions. Our findings underscore the significance of the tensor parametrization of matrix sensing, in combination with first-order methods, in achieving global optimality in such problems.

1 Introduction

This paper is dedicated to addressing the non-convex problem of matrix sensing, which has numerous practical applications and is rich in theoretical implications. Its canonical form can be written as:

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

$\mathcal{A}(\cdot) : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ is a linear operating consisting of m sensing matrices $\{A_i\}_{i=1}^m \in \mathbb{R}^{n\times n}$ where $\mathcal{A}(M) = [\langle A_1, M\rangle, \ldots, \langle A_m, M\rangle]^T$. The sensing matrices and the measurements $b = \mathcal{A}(M^*)$ are given, while M^* is an unknown low-rank matrix to be recovered from the measurements. The true rank of M^* is bounded by r, usually much smaller than the problem size n. More importantly, since $\mathcal{A}$ is linear, one can replace A_i with $(A_i + A_i^\top)/2$ without changing b, and therefore all sensing matrices can be assumed to be symmetric.

The aforementioned problem serves as an extension of both compressed sensing [1], which is widely applied in the field of medical imaging, and matrix completion [2, 3], which possesses an array of notable applications [4]. Additionally, this problem emerges in a variety of real-world situations such as phase retrieval [5–7], motion detection [8], and power system state estimation [9, 10]. A recent study by [11] established that any polynomial optimization problem can be converted into a series of problems following the structure of (1), thereby underscoring the significance of investigating this specific non-convex formulation. Within the realm of contemporary machine learning, (1) holds relevance as it is equivalent to the training problem for a two-layer neural network with quadratic activations [12]. In this context, m denotes the number of training samples, r is the size of the hidden layer, and the sensing matrices $A_i = x_i x_i^\top$ are rank-1, with x_i representing the i^{th} datapoint.

To solve (1), an increasingly popular approach is the Burer-Monteiro (BM) factorization [13], in which the low-rank matrix M is factorized into $M = XX^\top$ with $X \in \mathbb{R}^{n\times r}$, thereby omitting the

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

constraint, making it amenable to simple first-order methods such as gradient descent (GD), while scaling with $\mathcal{O}(nr)$ instead of $\mathcal{O}(n^2)$. The formulation can be formally stated as follows:

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

with $Z\in\mathbb{R}^{n\times r}$ being any ground truth representation such that $M^*=ZZ^\top$. Since (2) is a non-convex problem, it can have spurious local minima[1], making it difficult to recover M^* in general. The pivotal concept in solving (1) and (2) to optimality is the notion of Restricted Isometry Property (RIP), which measures the proximity between $\|\mathcal{A}(M)\|_F^2$ and $\|M\|_F^2$ for all low-rank matrices M. This proximity is captured by a constant δ_p, where $\delta_p=0$ means $\mathcal{A}(M)=M$ for matrices up to rank p, leading to exact isometry case, and $\delta_p\to 1$ implying a problematic scenario in which the proximity error is large. For a precise definition, please refer to Appendix A.3.

Conventional wisdom suggests that there is a sharp bound on the RIP constant that controls the recoverability of M^*, with $1/2$ being the bound for (2). [14, 15] prove that if $\delta_{2r}<1/2$, then all local minimizers are global minimizers, and conversely if $\delta_{2r}\geq 1/2$, counterexamples can be easily established. Similar bounds of $1/3$ are also derived for general objectives [16, 17], demonstrating the importance of the notion of RIP. However, more recent studies reveal that the technique of over-parametrization (by using $X\in\mathbb{R}^{n\times r_{\text{search}}}$ instead, with $r_{\text{search}}>r$) can take the sharp RIP bound to higher values [18, 19]. Recently, it has also been shown that using a semidefinite programming (SDP) formulation (convex relaxation) can lead to guaranteed recovery with a larger RIP bound that approaches 1 in the transition to the high-rank regime when $n\approx 2r$ [20]. These works all show the efficacy of over-parametrization, shedding light on a powerful way to find the global solution of complex non-convex problems. However, all of these techniques fail to handle real-world cases with $\delta_{2r}\to 1$ in the low-rank regime. To this end, a recent work [21] drew on important concepts from the celebrated Lasserre's Hierarchy [22] and proposed a lifted framework based on tensor optimization that could convert spurious local minimizers of (2) into strict saddle points in the lifted space, for arbitrary RIP constants in the $r=1$ case. We state this lifted problem below:

$$\min_{\mathbf{w}\in\mathbb{R}^{n\circ l}} \quad \|\langle \mathbf{A}^{\otimes l}, \mathbf{w}\otimes\mathbf{w}\rangle - b^{\otimes l}\|_F^2 \quad \text{(Lifted Problem, } r=1) \tag{3}$$

where $\mathbf{w}$ is an l-way, n-dimensional tensor, and $\mathbf{A}^{\otimes l}$ and $b^{\otimes l}$ are tensors "lifted" from $\mathcal{A}$ and b via tensor outer product. We defer the precise definition of tensors and their products to Section 2. The main theorem of [21] states that when $r=1$, for some appropriate l, the first-order points (FOP) of (2) will be converted to FOPs of (3) via lifting, and that spurious second order points (SOP) of (2) will be converted into strict saddles, under some technical conditions, provided that $\mathbf{w}$ is symmetric, and rank-1. This rank-1 constraint on the decision variable $\mathbf{w}$ is non-trivial, since finding the dominant rank-1 component of symmetric tensors is itself a non-convex problem in general, and requires a number of assumptions for it to be provably correct [23, 24]. This does not even account for the difficulties of maintaining the symmetric properties of tensors, which also has no natural guarantees. Therefore, although this lifted formulation may be promising in the pursuit of global minimum, there are still major questions to be answered. Most importantly, it is desirable to know *whether the symmetric, rank-1 condition is necessary, and if so, how to achieve it without explicit constraints?*

The necessity of the condition in question can be better understood through insights from [25]. The authors argue that over-parametrizing non-convex optimization problems can reshape the optimization landscape, with the effect being largely independent of the cost function and primarily determined by the parametrization. This notion is consistent with [21], which contends that over-parametrizing vectors into tensors can transform spurious local solutions into strict saddles. However, [25] specifically examines the parametrization from vectors/matrices to tensors, concluding that stationary points are not generally preserved under tensor parametrization, contradicting [21]. This implies that the symmetric, rank-1 constraints required in (3) are crucial for the conversion of spurious points.

It is essential to devise a method to encourage tensors to be near rank-1, with implicit regularization as a potential solution. There has been a recent surge in examining the implicit regularization effects in first-order optimization methods, such as gradient descent (GD) and stochastic gradient descent (SGD) [26], which has been well-studied in matrix sensing settings [27–29, 12]. This intriguing observation has prompted us to explore the possible presence of similar implicit regularization in tensor spaces. Our findings indicate that when applying GD to the tensor optimization problem

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

(3), an implicit bias can be detected with sufficiently small initialization points. This finding does not directly extend from its matrix counterparts due to the intricate structures of tensors, resulting in a scarcity of useful identities and well-defined concepts for even fundamental properties such as eigenvalues. Furthermore, we show that when initialized at a symmetric tensor, the entire GD trajectory remains symmetric, completing the requirements.

In this paper, we demonstrate that over-parametrization alone does not inherently simplify non-convex problems. However, employing a suitable optimization algorithm offers a remarkably straightforward solution, as this specific algorithm implicitly constrains our search to occur within a compact representation of the over-parametrized space without necessitating manual embeddings or transformations. This insight further encourages the investigation of a (parametrization, algorithm) pair for solving non-convex problems, thereby enhancing our understanding of achieving global optimality in non-convex problems.

1.1 Related Works

Over-parametrization in matrix sensing. Except for the lifting formulation (3), there are two mainstream approaches to over-parametrization in matrix sensing. The first one is done via searching over $Y \in \mathbb{R}^{n\times r_{\text{search}}}$ instead of $X \in \mathbb{R}^{n\times r}$, and using some distance metric to minimize the distance between $\mathcal{A}(YY^\top)$ and b. Using an l_2 norm, [18, 19] established that if $r_{\text{search}} > r[(1+\delta_n)/(1-\delta_n) - 1]^2/4$, with $r \leq r_{\text{search}} < n$, then every second-order point $\hat{Y} \in \mathbb{R}^{n\times r_{\text{search}}}$ satisfies that $\hat{Y}\hat{Y}^\top = M^*$. [30] showed that even in the over-parametrized regime, noise can only finitely influence the optimization landscape. [29] offered similar results for an l_1 loss under good enough RIP constant. Another popular approach to over-paramerization is to use a convex SDP formulation, which is a convex relaxation of (1) [31]. It has been known for years that as long as $\delta_{2r} < 1/2$, then the global optimality of the SDP formulation correspond to the ground truth M^* [32]. Recently [20] updated this bound to $2r/(n + (n-2r)(2l-5))$, which can approach 1 if $n \approx 2r$.

Algorithm regularization in over-parametrized matrix sensing. [12, 33] prove that the convergence to global solution via GD is agnostic of r_{search}, in that it only depends on initialization scale, step-size, and RIP property. [29] demonstrates the same effect for an l_1 norm, and further showed that a small initialization nullifies the effect of over-parametrization. Besides these works, [27] refined this analysis, showing that via a sufficiently small initialization, the GD trajectory will make the solution implicitly penalize towards rank-r matrices after a small number of steps. [28] took it even further by showing that the GD trajectory will first make the matrix rank-1, rank-2, all the way to rank-r, in a sequential way, thereby resembling incremental learning.

Implicit bias in tensor learning. The line of work [34–36] demonstrates that for a class of tensor factorization problems, as long as the initialization scale is small, the learned tensor via GD will be approximately rank-1 after an appropriate number of steps. Our paper differs from this line of work in three meaningful ways: 1) The problem considered in those works are optimization problems over vectors, not tensors, and therefore the goal is to learn the structure of a known tensor, rather than learning a tensor itself; 2) Our proof relies directly on tensor algebra instead of adopting a dynamical systems perspective, providing deeper insights into tensor training dynamics while dispensing with the impractical assumption of an infinitesimal step-size.

1.2 Main Contributions

1. We demonstrate that, beyond vector and matrix learning problems, optimization of differentiable objectives, such as the l_2 norm, through Gradient Descent (GD) can encourage a more compact representation for tensors as decision variables. This results in tensors being approximately rank-1 after a number of gradient steps. To achieve this, we employ an innovative proof technique grounded in tensor algebra and introduce a novel tensor eigenvalue concept, the variational eigenvalue (v-eigenvalue), which may hold independent significance due to its ease of use in optimization contexts.
2. We show that if a tensor is a first-order point of the lifted objective (3) and is approximately rank-1, then its rank-1 component can be mapped to an FOP of (2), implying that all FOPs of (3) lie in a small sphere around the lifted FOPs of (2). Furthermore, these FOPs possess an escape direction when reasonably distant from the ground truth solution, irrespective of the Restricted Isometry Property (RIP) constants.

3. We present a novel lifted framework that optimizes over symmetric tensors to accommodate the over-parametrization of matrix sensing problems with arbitrary r. This approach is necessary because directly extending the work of [21] from $r = 1$ to higher values may lead to non-cubical and, consequently, non-symmetric tensors.

2 Preliminaries

Please refer to Appendix A.1 and A.2 for the notations and definitions of first-order and second-order conditions. Here, we introduce two concepts that are critical in understanding our main results.

Definition 1 (Tensors and Products)**.** We define an l-way tensor as:

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

Moreover, if $n_1 = \dots = n_l$, then we call this tensor an l-order (or l-way), n-dimensional tensor. $\mathbb{R}^{n \circ l}$ is an abbreviated notion for $n \circ l \coloneqq n \times \dots \times n$. In this work, tensors are denoted with bold letters unless specified otherwise. The tensor outer product, denoted as $\otimes$, of 2 tensors $\mathbf{a}$ and $\mathbf{b}$, respectively of orders l and p, is a tensor of order $l + p$, namely $\mathbf{c} = \mathbf{a} \otimes \mathbf{b}$ with $c_{i_1 \dots i_l j_1 \dots j_p} = a_{i_1 \dots i_l} b_{j_1 \dots j_p}$. We also use the shorthand $\mathbf{a}^{\otimes l}$ for repeated outer product of l times for arbitrary tensor/matrix/vector $\mathbf{a}$. $\langle \mathbf{a}, \mathbf{b} \rangle_{i_1, \dots, i_d}$ denotes tensor inner product along dimensions $i_1, \dots, i_d$ (with respect to the first tensor), in which we simply sum over the specified dimensions after the outer product $\mathbf{a} \otimes \mathbf{b}$ is calculated. This means that the inner product is of $l + p - 2d$ orders. Please refer to Appendix A.3 for a more in-depth review on tensors, especially on its symmetry and rank.

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

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

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

3 The Lifted Formulation for General r

A natural extension of (3) to general r requires that instead of optimizing over $X \in \mathbb{R}^{n \times r}$, we optimize over $\mathbb{R}^{[n \times r] \circ l}$ tensors, and simply making tensor outer products between $\mathbf{w}$ to be inner products. However, such a tensor space is non-cubical, and subsequently not symmetric. This is the higher-dimensional analogy of non-square matrices, which lacks a number of desirable properties, as per the matrix scenario. In particular, it is necessary for our approach to optimize over a cubical, symmetric tensor space since in the next section we prove that there exists an implicit bias of the gradient descent algorithm under that setting.

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

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

Such $\mathbf{P}$ can be easily constructed via filling appropriate scalar "1"s in the tensor. Via Lemma A.4, we also know that

$$\langle \mathbf{P}^{\otimes l}, \mathrm{vec}(X)^{\otimes l} \rangle_{3*[l]} = (\langle \mathbf{P}, \mathrm{vec}(X) \rangle_3)^{\otimes l} = X^{\otimes l} \tag{4}$$

where $[l]$ denotes the integer set $[1, \dots, l]$, and $c * [l]$ denotes $[c, 2c, \dots, c * l]$ for some $c \in \mathbb{Z}^+$. For notational convenience, we abbreviate $\langle \mathbf{P}^{\otimes l}, \mathbf{w} \rangle_{3*[l]}$ as $\mathbf{P}(\mathbf{w})$ for any arbitrary z-dimensional tensor $\mathbf{w}$ where z can be broken down into the product of two positive integers. Thus, using (4), we can extend (3) to a problem of general r, yet still defined over a cubical tensor space:

$$\min_{\mathbf{w} \in \mathbb{R}^{nr \circ l}} \quad \|\langle \mathbf{A}^{\otimes l}, \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{w}) \rangle_{2*[l]} \rangle - b^{\otimes l}\|_F^2 \quad \text{(Lifted formulation, general } r\text{)} \tag{5}$$

Let us define a 3-way tensor $\mathbf{A} \in \mathbb{R}^{m\times n\times n}$ so that $\mathbf{A}_{kij} = (A_k)_{ij}\ \forall k \in [m], (i,j) \in [n]\times[n]$. Define $f^l(\cdot) : \mathbb{R}^{n\circ 2l} \mapsto \mathbb{R}$ and $h^l(\cdot) : \mathbb{R}^{[n\times r]\circ l} \mapsto \mathbb{R}$ as $f^l(\mathbf{M}) \coloneqq \|\langle \mathbf{A}^{\otimes l}, \mathbf{M}\rangle - b^{\otimes l}\|_F^2$ and $h^l(\mathbf{w}) = f^l(\langle \mathbf{w}, \mathbf{w}\rangle_{2*[l]})$, with $\nabla f^l(\cdot) = \nabla_{\mathbf{M}} f^l(\cdot)$ and $\nabla h^l(\cdot) = \nabla_{\mathbf{w}} h^l(\cdot)$.

We prove that (5) has all the good properties detailed in [21] for (3). In particular, we prove that the symmetric, rank-1 FOPs of (5) have a one-to-one correspondence with those of (2), and that those FOPs that are reasonably separated from M^* or have a small r^{th} singular value can be converted to strict saddle points via some level of lifting. For the detailed theorems and proofs, please refer to Appendix B.

4 Implicit Bias of Gradient Descent in Tensor Space

In this section, we study why and how applying gradient descent to (5) will result in an implicit bias towards to rank-1 tensors. Prior to presenting the proofs, we shall elucidate the primary intuition behind how GD contributes to the implicit regularization of (2). This will aid in comprehending the impact of implicit bias on (5), as they share several crucial observations, albeit encountering greater technical hurdles. Consider the first gradient step of (2), initialized at a random point $X_0 \in \mathbb{R}^{n\times r_{\text{search}}} = \epsilon X$ with $\|X\|_F^2 = 1$ and $r_{\text{search}} \geq r$:

$$\begin{aligned} X_1 &= X_0 - \eta\nabla h(X_0) = \left(I + \eta\left[\mathcal{A}^*\mathcal{A}(M^*)\right]\right) X_0 - \left[\mathcal{A}^*\mathcal{A}(X_0X_0^\top)\right] X_0 \\ &= \left(I + \eta\left[\mathcal{A}^*\mathcal{A}(M^*)\right]\right) X_0 - \epsilon^2\left[\mathcal{A}^*\mathcal{A}(XX^\top)\right] X_0 \\ &= \left(I + \eta\left[\mathcal{A}^*\mathcal{A}(M^*)\right]\right) X_0 + \mathcal{O}(\epsilon^3) \end{aligned}$$

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

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

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

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

However, when tackling the implicit regularization problem in tensor space, one key deviation from the aforementioned procedure is that L_s/α_s will be relatively large, as otherwise there will be no spurious solutions, even in the noisy case [14, 30], which is also the motivation for using a lifted framework in the first place. Therefore, instead of saying that $\mathcal{A}^*\mathcal{A}(M^*) \approx M^*$, we aim to show that the gap between the eigenvalues of a comparable tensor term will enlarge as we increase l, making the tensor predominantly rank-1. This observation demonstrates the power of the lifting technique, while at the same time eliminates the critical dependence on a small L_s/α_s factor that is in practice often unachievable due to requiring sample numbers m in the asymptotic regime [37].

Therefore, in order to establish an implicit regularization result for (5), there are four major steps that need to be taken:

1. Proving that a point on the GD trajectory $\mathbf{w}_t$ admits a certain breakdown in the form $\mathbf{w}_t = \langle \mathbf{Z}_t, \mathbf{w}_0\rangle - \mathbf{E}_t$ for some $\mathbf{Z}_t$ and $\mathbf{E}_t$.
2. Proving that the spectral norm (equivalence of largest singular value) of $\mathbf{E}_t$ is small (scales with initialization scale ϵ)
3. Proving that $\langle \mathbf{Z}_t, \mathbf{w}_0\rangle$ has a large separation between its largest and second largest eigenvalues using a tensor version of Weyl's inequality.
4. Showing that, with the above holding true, $\mathbf{w}_t$ is predominantly rank-1 after some step t_*.

Lemmas 12, 13, 2, and Theorem 1 correspond to the above four steps, respectively. The reader is referred to the lemmas and theorem for more details.

4.1 A Primer on Tensor Algebra and Maintaining Symmetric Property

We start with the spectral norm of tensors, which resembles the operator norm of matrices [38].

Definition 3. Given a cubic tensor $\mathbf{w} \in \mathbb{R}^{n \circ l}$, its spectral norm $\|\cdot\|_S$ is defined respectively as:

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

There are many definitions for tensor eigenvalues [39], and in this paper we introduce a novel variational characterization of eigenvalues that resembles the Courant-Fisher minimax definition for eigenvalues of matrices, called the v-Eigenvalue. We denote the i^{th} v-Eigenvalue of $\mathbf{w}$ as $\lambda_i^v(\mathbf{w})$. Note this is a new definition that is first introduced in this paper and might be of independent interest outside of the current scope.

Definition 4 (Variational Eigenvalue of Tensors)**.** For a given tensor $\mathbf{w} \in \mathbb{R}^{n \circ l}$, we define its k^{th} variational eigenvalue (v-Eigenvalue) $\lambda_k^v(\mathbf{w})$ as

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

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

Next, since most of our analysis relies on the symmetry of the underlying tensor, it is desirable to show that every tensor along the optimization trajectory of GD on (5) remains symmetric if started from a symmetric tensor. Please find its proof in Appendix C.2.

Lemma 1. *If the GD trajectory of* (5) $\{\mathbf{w}_t\}_{t=0}^{\infty}$ *is initialized at a symmetric rank-1 tensor* $\mathbf{w}_0$*, then* $\{\mathbf{w}_t\}_{t=0}^{\infty}$ *will all be symmetric.*

4.2 Main Ideas and Proof Sketch

In this subsection, we highlight the main ideas behind implicit bias in GD. Lemma 12 and 13 details the first and second step, and are deferred to Appendix C.2. The proofs to the results of this section can also be found in that appendix. The lemmas alongside with their proofs are highly technical and not particularly enlightening, therefore omitted here for simplicity. However, the most important takeaway is that for the t^{th} iterate along the GD trajectory of (5), we have the decomposition

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

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

Lemma 2. *Given* $\mathbf{w}_t$ *along the GD trajectory of* (5)*, its first two v-eigenvalues, as defined in definition 4, satisfy the relation*

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \leq \frac{\|x_0\|_2^l (1+\eta\sigma_2^l(U))^t + \|\mathbf{E}_t\|_S/\epsilon}{|v_1^\top x_0|^l (1+\eta\sigma_1^l(U))^t - \|\mathbf{E}_t\|_S/\epsilon} = \frac{\|x_0\|_2^l (1+\eta\sigma_2^l(U))^t + \mathcal{O}(\epsilon^2)}{|v_1^\top x_0|^l (1+\eta\sigma_1^l(U))^t - \mathcal{O}(\epsilon^2)} \tag{6}$$

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

Lemma 2 showcases that when ϵ is small, the ratio between the largest and second largest v-eigenvalues of $\mathbf{w}$ is dominated by $(\|x_0\|_2^l (1+\eta\sigma_2^l(U))^t)/(|v_1^\top x_0|^l (1+\eta\sigma_1^l(U))^t)$.

Now, if either $\|x_0\|_2^l$ is large or $|v_1^\top x_0|^l$ approaches 0 in value, then the ratio may be relatively large, contradicting our claim. However, this issue can be easily addressed by letting $x_0 = v_1 + g \in \mathbb{R}^{nr}$, where g is a vector with each entry being i.i.d sampled from the Gaussian distribution $\mathcal{N}(0, \rho)$. Note that since $U = \langle \mathbf{A}_r, b \rangle_3$, we can calculate U and v_1 directly. Lemma 14 in Appendix C.2 shows that with this initialization, $|v_1^\top x_0|^l = \mathcal{O}(1)$ and $\|x_0\|_2^l = \mathcal{O}(1)$ with high probability if we select $\rho = \mathcal{O}(1/nr)$. Therefore, the t^{th} iterate along the GD trajectory of (5) satisfies

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \asymp \frac{(1+\eta\sigma_2^l(U))^t}{(1+\eta\sigma_1^l(U))^t} \tag{7}$$

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

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

Theorem 1. *Given the optimization problem* (5) *and its GD trajectory over some finite horizon T, i.e., $\{\mathbf{w}_t\}_{t=0}^T$ with $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla h^l(\mathbf{w}_t)$, where η is the stepsize, then there exist $t(\kappa, l) \geq 1$ and $\kappa < 1$ such that*

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \leq \kappa, \qquad \forall t \in [t(\kappa, l), t_T] \tag{8}$$

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

$$t(\kappa, l) = \left\lceil \ln\left(\frac{\|x_0\|_2^l}{\kappa|v_1^\top x_0|^l}\right) \ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1} \right\rceil \tag{9}$$

By using the initialization introduced in Lemma 14, we can improve the result of Theoerem 1, which does not need ϵ to be arbitrarily small. The full details are presented in Corollary 1 in Appendix C.2, stating that as along as $t \asymp \ln(1/\kappa)\ln\left((1+\eta\sigma_1^l(U))/(1+\eta\sigma_2^l(U))\right)^{-1}$, $\mathbf{w}_t$ will be κ-rank-1, as long as ϵ is chosen as a function of U, r, n, L_s, and κ. Note that we say a tensor $\mathbf{w}$ is "κ-rank-1" if $\lambda_2^v(\mathbf{w})/\lambda_1^v(\mathbf{w}) \leq \kappa$.

5 Approximate Rank-1 Tensors are Benign

Now that we have established the fact that performing gradient descent on (5) will penalize the tensor towards rank-1, it begs the question whether approximate rank-1 tensors can also escape from saddle points, which is the most important question under study in this paper. Please find the proofs to the results in this section in Appendix D.

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

Proposition 1. *Given a symmetric tensor $\mathbf{w} \in \mathbb{R}^{n\times\cdots\times n}$, it can be decomposed into two terms, namely a term consisting of its dominant component and another term that is orthogonal to this direction:*

$$\mathbf{w} = \pm\lambda_1^v(\mathbf{w}) w_s^{\otimes l} + \mathbf{w}^\dagger := \mathbf{w}_\sigma + \mathbf{w}^\dagger, \quad w_s \in \mathbb{R}^n,\ \|w_s\|_2 = 1 \tag{10}$$

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

Next, we characterize the first-order points of (5) with approximate rank-1 tensors in mind. Previously, we showed that if a given FOP of (5) is symmetric and rank-1, it has a one-to-one correspondence with FOPs of (2). However, if the FOPs of (5) are not exactly rank-1, but instead κ-rank-1, it is essential to understand whether they maintain the previous properties. This will be addressed below.

Proposition 2. *Assume that a symmetric tensor $\mathbf{w} \in \mathbb{R}^{n\times\cdots\times n}$ is an FOP of* (5)*, meaning that* (17a) *holds. If it is a κ-rank-1 tensor with $\kappa \leq \mathcal{O}(1/\|M^*\|_F^2)$, then it admits a decomposition as*

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

with $\mathrm{mat}(\hat{w}) \in \mathbb{R}^{n\times r}$ being an FOP of (2) *and $\|\mathbf{w}^\dagger\|_S \leq \kappa\lambda_1^v(\mathbf{w})$ by definition.*

The proposition above asserts that for any given FOP of (5), if it is κ-rank-1 rather than being truly rank-1, it will consist of a rank-1 term representing a lifted version of an unlifted FOP, as well as a term with a small spectral norm. Referring to (58), it is possible to achieve a significantly low κ through a moderate number of iterations. This result, considered the cornerstone of this paper, demonstrates that the use of gradient descent with small initialization will find critical points that are lifted FOPs of (2) with added noise, maintaining a robust association between FOPs of (5) and (2). This finding also facilitates this subsequent theorem:

Theorem 2. *Assume that a symmetric tensor $\hat{\mathbf{w}} \in \mathbb{R}^{nr\circ l}$ is an FOP of* (5) *that is κ-rank-1 with $\kappa \leq \mathcal{O}(1/\|M^*\|_F^2)$. Consider its major spectral decomposition $\hat{\mathbf{w}} = \lambda_S \hat{x}^{\otimes l} + \hat{\mathbf{w}}^\dagger$ with $\hat{x} \in \mathbb{R}^{nr}$, then it has a rank-1 escape direction if $\hat{X} = \mathrm{mat}(\hat{x})$ satisfies the inequality*

$$\|M^* - \hat{X}\hat{X}^\top\|_F^2 \geq \frac{L_s}{\alpha_s}\lambda_r(\hat{X}\hat{X}^\top)\,\mathrm{tr}(M^*) + \mathcal{O}(r\kappa^{1/l}) \tag{11}$$

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

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

This theorem conveys the message that by running GD on (5), all critical points have escape directions as long as the point is not close to the ground truth solution. In Appendix B, we present Theorem 7 to provide sufficient conditions for the conversion to hold globally when (11) is hard to hold.

6 Numerical Experiments

In this section[2], after we run a given algorithm on (5) to completion and obtain a final tensor $\mathbf{w}_T$, we then apply tensor PCA (detailed in Appendix F) on $\mathbf{w}_T$ to extract its dominant rank-1 component and recover $X_T \in \mathbb{R}^{n\times r}$ such that $(\mathbf{w}_T)_s = \lambda_s\,\mathrm{vec}(X_T)^{\otimes l}$. Since $\mathbf{w}_T$ will be approximately rank-1, the success of this operation is expected [23, 24]. We consider a trial to be successful if the recovered X_T satisfies $\|X_T X_T^\top - M^*\|_F \leq 0.05$. We also initialize our algorithm as per Lemma 14.

6.1 Perturbed Matrix Completion

The perturbed matrix completion problem is introduced in [20], which is a noisy version of classic matrix completion problems. The $\mathcal{A}$ operator is introduced as

$$\mathcal{A}_\rho(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij}, & \text{if } (i,j) \in \Omega \\ \rho\mathbf{M}_{ij}, & \text{otherwise} \end{cases}, \tag{12}$$

where Ω is a measurement set such that $\Omega = \{(i,i),(i,2k),(2k,i)|\ \forall i \in [n], k \in [\lfloor n/2 \rfloor]\}$. [20] has proved that each such instance has $\mathcal{O}(2^{\lceil n/2\rceil} - 2)$ spurious local minima, while it satisfies the RIP property with $\delta_{2r} = (1-\rho)/(1+\rho)$ for some sufficiently small ρ. This implies that common first-order methods fail with high probability for this class of problems. In our experiment, we apply both lifted and unlifted formulations to (12) with $\rho = 0.01$, yielding $\delta_{2r} \approx 1$. We test different values of n and ϵ, using a lifted level of $l = 3$. We ran 10 trials each to calculate success rate. If unspecified in the plot, we default $n = 10$, $\epsilon = 10^{-7}$. Figure 1 reveals a higher success rate for the lifted formulation across different problem sizes, with smaller problems performing better as expected (since larger problems require a higher lifting level). Success rates improve with smaller ϵ, emphasizing the importance of small initialization. We employed customGD, a modified gradient descent algorithm with heuristic saddle escaping. This algorithm will deterministically escape from critical points utilizing knowledge from the proof of Theorem 4. For details please refer to Appendix F. Furthermore, to showcase the implicit penalization affects of GD, we obtained approximate measures for $\lambda_2^v(\mathbf{w}_t)/\lambda_1^v(\mathbf{w}_t)$ (since exactly solving for them is NP-hard) along the trajectory, and presented the results and methods in Appendix E.

Additionally, we examine different algorithms for (5), including customGD, vanilla GD, perturbed GD ([40], for its ability to escape saddles), and ADAM [41]. Figure 2 suggest that ADAM is an effective optimizer with a high success rate and rapid convergence, indicating that momentum acceleration may not hinder implicit regularization and warrants further research. Perturbed GD performed poorly, possibly due to random noise disrupting rank-1 penalization.

6.2 Shallow Neural Network Training with Quadratic Activation

It has long been known that the matrix sensing problem (2) includes the training of two-layer neural networks (NN) with quadratic activation as a special case [12]. In summary, the output of the neural

[2]'https://github.com/anonpapersbm/implicit_bias_tensor',run on 2021 Macbook Pro

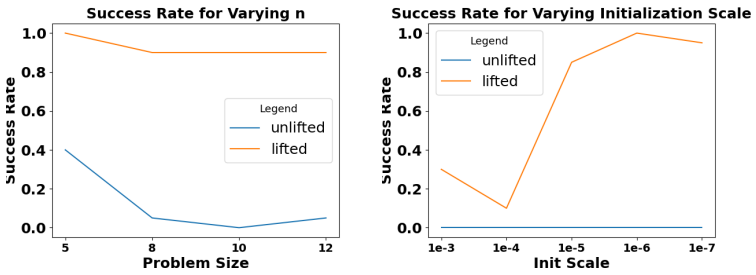

Figure 1: Success rate of the lifted formulation versus the unlifted formulation against varying n and ϵ. The algorithm of choice is CustomGD (details in Appendix F).

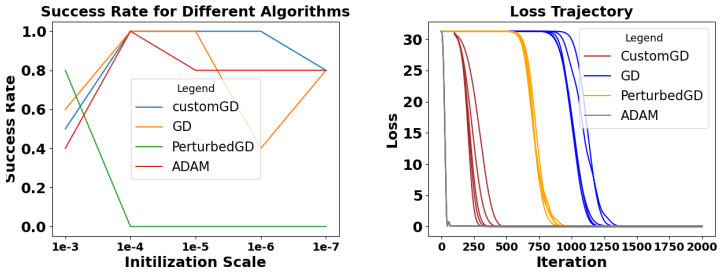

Figure 2: Performance of different algorithms applied to the lifted formulation (5).

network $y \in \mathbb{R}^m$ with respect to m inputs $\{d_i\}_{i=1}^m \in \mathbb{R}^n$ can be expressed as $y_i = \mathbf{1}^\top q(X^\top d_i)$, which implies $y_i = \langle d_i d_i^\top, XX^\top \rangle$, where $q(\cdot)$ is the element-wise quadratic function and $X \in \mathbb{R}^{n \times r}$ in (2) represents the weights of the neural network. Thus r represents the number of hidden neurons. In our experiment, we demonstrate that when m is small, the lifted framework (5) outperforms standard neural network training in success rate, yielding improved recovery of the true weights. We set the hidden neurons number to be n for the standard network training, thereby comparing the existing over-parametrization framework (recall Section 1.1, with $r_{\text{search}} = n$) with the lifted one . We employ the ADAM optimizer for both methods. Table 2 showcases the success rate under various problem and sample sizes. Sampling both data and true weights $Z \in \mathbb{R}^{n \times r}$ from an i.i.d Gaussian distribution, we calculate the observations y and attempt to recover Z using both approaches. As the number of samples increases, so does the success rate, with the lifted approach offering significantly better accuracy overall, even when the standard training has a 0% success rate.

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

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

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

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

Table 2: Success rate of NN training using (5) and original formulation. The number inside the parentheses denotes the success rate of the original formulations. $\epsilon = 10^{-5}$ and $l = 3$.

7 Conclusion

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

8 Acknowledgement

This work was supported by grants from ARO, ONR, AFOSR, NSF, and the UC Noyce Initiative.

References

[1] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[3] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[4] L. T. Nguyen, J. Kim, and B. Shim, "Low-rank matrix completion: A contemporary survey," *IEEE Access*, vol. 7, pp. 94215–94237, 2019.

[5] A. Singer, "Angular synchronization by eigenvectors and semidefinite programming," *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 20–36, 2011.

[6] N. Boumal, "Nonconvex phase synchronization," *SIAM Journal on Optimization*, vol. 26, no. 4, pp. 2355–2377, 2016.

[7] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: A contemporary overview," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 87–109, 2015.

[8] S. Fattahi and S. Sojoudi, "Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis," *Journal of Machine Learning Research*, vol. 21, pp. 1–51, 2020.

[9] Y. Zhang, R. Madani, and J. Lavaei, "Conic relaxations for power system state estimation with line measurements," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1193–1205, 2017.

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

[11] I. Molybog, R. Madani, and J. Lavaei, "Conic optimization for quadratic regression under sparse noise," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 7994–8029, 2020.

[12] Y. Li, T. Ma, and H. Zhang, "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations," in *Conference On Learning Theory*, pp. 2–47, PMLR, 2018.

[13] S. Burer and R. D. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming*, vol. 95, no. 2, pp. 329–357, 2003.

[14] R. Y. Zhang, S. Sojoudi, and J. Lavaei, "Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery," *Journal of Machine Learning Research*, vol. 20, no. 114, pp. 1–34, 2019.

[15] Z. Ma, Y. Bi, J. Lavaei, and S. Sojoudi, "Sharp restricted isometry property bounds for low-rank matrix recovery problems with corrupted measurements," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 7672–7681, 2022.

[16] W. Ha, H. Liu, and R. F. Barber, "An equivalence between critical points for rank constraints versus low-rank factorizations," *SIAM Journal on Optimization*, vol. 30, no. 4, pp. 2927–2955, 2020.

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

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

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

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

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

[22] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM Journal on optimization*, vol. 11, no. 3, pp. 796–817, 2001.

[23] E. Kofidis and P. A. Regalia, "On the best rank-1 approximation of higher-order supersymmetric tensors," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 3, pp. 863–884, 2002.

[24] L. Wu, X. Liu, and Z. Wen, "Symmetric rank-1 approximation of symmetric high-order tensors," *Optimization Methods and Software*, vol. 35, no. 2, pp. 416–438, 2020.

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

[26] P. Li, X. Liang, and H. Song, "A survey on implicit bias of gradient descent," in *2022 14th International Conference on Computer Research and Development (ICCRD)*, pp. 108–114, IEEE, 2022.

[27] D. Stöger and M. Soltanolkotabi, "Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23831–23843, 2021.

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

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

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

[31] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[32] T. T. Cai and A. Zhang, "Sharp rip bound for sparse signal and low-rank matrix recovery," *Applied and Computational Harmonic Analysis*, vol. 35, no. 1, pp. 74–93, 2013.

[33] J. Zhuo, J. Kwon, N. Ho, and C. Caramanis, "On the computational and statistical complexity of over-parameterized matrix sensing," *arXiv preprint arXiv:2102.02756*, 2021.

[34] N. Razin, A. Maman, and N. Cohen, "Implicit regularization in tensor factorization," in *International Conference on Machine Learning*, pp. 8913–8924, PMLR, 2021.

[35] N. Razin, A. Maman, and N. Cohen, "Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks," in *International Conference on Machine Learning*, pp. 18422–18462, PMLR, 2022.

[36] R. Ge, Y. Ren, X. Wang, and M. Zhou, "Understanding deflation process in over-parametrized tensor decomposition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1299–1311, 2021.

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

[38] L. Qi, S. Hu, X. Zhang, and Y. Chen, "Tensor norm, cubic power and gelfand limit," *arXiv preprint arXiv:1909.10942*, 2019.

[39] L. Qi, "The spectral theory of tensors (rough version)," *arXiv preprint arXiv:1201.3424*, 2012.

[40] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, "On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points," *Journal of the ACM (JACM)*, vol. 68, no. 2, pp. 1–29, 2021.

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

[42] Q. Li, Z. Zhu, and G. Tang, "The non-convex geometry of low-rank matrix optimization," *Information and Inference: A Journal of the IMA*, vol. 8, no. 1, pp. 51–96, 2019.

[43] T. G. Kolda, "Numerical optimization for symmetric tensor decomposition," *Mathematical Programming*, vol. 151, no. 1, pp. 225–248, 2015.

[44] K. B. Petersen, M. S. Pedersen, *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, no. 15, p. 510, 2008.

[45] Z. Ma and S. Sojoudi, "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.

[46] G. Zhang and R. Y. Zhang, "How many samples is a good initial point worth in low-rank matrix recovery?," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 12583–12592, 2020.

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

[48] L.-H. Lim and P. Comon, "Blind multilinear identification," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1260–1280, 2013.

[49] P. Comon, G. Golub, L.-H. Lim, and B. Mourrain, "Symmetric tensors and symmetric tensor rank," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1254–1279, 2008.

[50] G. Ni, "Hermitian tensor and quantum mixed state," *arXiv preprint arXiv:1902.02640*, 2019.

[51] S. Y. Chang, "Hanson-wright inequality for random tensors under einstein product," *arXiv preprint arXiv:2111.12169*, 2021.

[52] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press, 2018.

A Additional Definitions and Supporting Lemmas

A.1 Notations

In this paper, $\sigma_i(M)$ 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. For a matrix M, $\mathrm{vec}(M)$ is the usual vectorization operation by stacking the columns of the matrix M into a vector. For a vector $v \in \mathbb{R}^{n^2}$, $\mathrm{mat}(v)$ converts v to a square matrix and $\mathrm{mat}_S(v)$ converts v to a symmetric matrix, i.e., $\mathrm{mat}(v) = M$ and $\mathrm{mat}_S(v) = (M + M^T)/2$, where $M \in \mathbb{R}^{n\times n}$ is the unique matrix satisfying $v = \mathrm{vec}(M)$. $[n]$ denotes the integer set $[1, \ldots, n]$, and $\circ l$ stands for the shorthand of repeated cartesian product $\times \cdots \times$ for l times. The symbol $\oslash$ denotes the kronecker product, while $\otimes$ denotes tensor outer product. $\asymp$ denotes "asymptotic to", meaning that the two terms on both sides of this symbol have the same order of magnitude.

A.2 Critical Conditions for Unlifted Problem

We present the FOP and SOP conditions for the unlifted problem as our benchmark.

Lemma 3. *The vector $\hat{X} \in \mathbb{R}^{n\times r}$ is an SOP of* (2) *if and only if*

$$\nabla f(\hat{X}\hat{X}^\top)\hat{X} = 0, \tag{13}$$

$$2\langle \nabla f(\hat{X}\hat{X}^\top), UU^\top\rangle + [\nabla^2 f(\hat{X}\hat{X}^\top)](\hat{X}U^\top + U\hat{X}^\top, \hat{X}U^\top + U\hat{X}^\top) \geq 0 \quad \forall U \in \mathbb{R}^{n\times r} \tag{14}$$

with (13) *being the necessary and sufficient condition for $\hat{X}$ to be an FOP.*

A proof to the above lemma can be found in many matrix sensing literatures, including [16, 17, 42], etc.

A.3 Additional Definitions

Definition 5 (RIP, [2])**.** 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 all matrices $M \in \mathbb{R}^{n\times n}$ satisfying $\mathrm{rank}(M) \leq p$.

Definition 6 (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:

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

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

Definition 7 (Rank of Tensors)**.** The rank of a cubic tensor $\mathbf{a} \in \mathbb{R}^{n\circ l}$ is defined as

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

for some vector $u_i, \ldots, w_i \in \mathbb{R}^n$. Furthermore, according to [43], if $\mathbf{a}$ is a symmetric tensor, then it can be decomposed as:

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

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

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

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

Definition 8 (Tensor Multiplication)**.** Outer product is an operation carried out on a pair of tensors, denoted as $\otimes$. The outer product of 2 tensors $\mathbf{a}$ and $\mathbf{b}$, respectively of orders l and p, is a tensor of order $l+p$, denoted as $\mathbf{c} = \mathbf{a} \otimes \mathbf{b}$ such that:

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

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

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

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 that $q = 1$ 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 $\langle \mathbf{a}, \mathbf{b} \rangle_{q,\dots,s}$.

A.4 Technical Lemmas

Lemma 4 (Section 10.2 [44])**.** *For four arbitrary matrices A, B, C, D of compatible dimensions, it holds that*

$$\langle A \otimes B, C \otimes D \rangle_{2,4} = AC \otimes BD \tag{15}$$

Lemma 5 ([45])**.** *For any SOP $\hat{X}$ of* (2)*, define G as $G \coloneqq -\lambda_{min}(\nabla f(\hat{X}\hat{X}^\top))$, and L_s be the RSS constant. Then it holds that*

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

where r is the search rank of (2)*.*

Lemma 6. *Given an FOP $\hat{X}$ of* (2)*, it holds that*

$$\lambda_r(\hat{X}\hat{X}^\top) < \sqrt{\frac{2L_s}{r\alpha_s}} \|M^*\|_F \tag{16}$$

Proof of Lemma 6. Lemma 6 of [17] states that given an arbitrary constant λ and matrix $X \in \mathbb{R}^{n \times r}$, one can write

$$\|XX^\top\|_F^2 \geq \max\left\{ \frac{2L_s}{\alpha_s} \|M^*\|_F^2, (\frac{2\lambda\sqrt{r}}{\alpha_s})^{4/3} \right\} \implies \|\nabla h(X)\|_F \geq \lambda$$

A simple negation to both sides gives

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

If we set $X = \hat{X}$, then left-hand side of the above inequality is automatically satisfied for small values of λ since $\|\nabla h(\hat{X})\|_F = 0$, and thus we conclude that

$$\|\hat{X}\hat{X}^\top\|_F^2 < \frac{2L_s}{\alpha_s} \|M^*\|_F^2$$

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

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

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

B Additional Details for Lifted Formulation of General r

We analyze (5) and generalize the results of [21] to $r > 1$. We start with the characterization of FOPs and SOPs of (5).

Lemma 7. *The tensor $\hat{\mathbf{w}} \in \mathbb{R}^{nr\circ l}$ is an SOP of* (5) *if and only if*

$$\langle \nabla f^l(\langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}})\rangle_{2*[l]}), \mathbf{P}(\hat{\mathbf{w}})\rangle_{2*[l]} = 0, \tag{17a}$$

$$\begin{aligned} 2\langle \nabla f^l(\langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}})\rangle_{2*[l]}), \langle \mathbf{P}(\Delta), \mathbf{P}(\Delta)\rangle_{2*[l]} + \\ \|\langle \mathbf{A}^{\otimes l}, \langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\Delta)\rangle_{2*[l]} + \langle \mathbf{P}(\Delta), \mathbf{P}(\hat{\mathbf{w}})\rangle_{2*[l]}\rangle\|_F^2 \geq 0 \quad \forall \Delta \in \mathbb{R}^{nr\circ l} \end{aligned} \tag{17b}$$

with (17b) *being a necessary and sufficient condition for $\hat{\mathbf{w}}$ to be a FOP.*

Proof of Lemma 7. We have

$$\nabla f^l(\mathbf{M}) = \langle \langle \mathbf{A}^{\otimes l}, \mathbf{M} - \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})\rangle, \mathbf{A}^{\otimes l}\rangle_{1,4,\ldots,3l-2} \tag{18}$$

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

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

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

$$D_{\mathbf{w}}\mathcal{M}(\mathbf{v}) = \langle \mathbf{P}(\mathbf{v}), \mathbf{P}(\mathbf{w})\rangle_{2*[l]} + \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{v})\rangle_{2*[l]}. \tag{19}$$

Combining (18) and (19) gives that

$$D_{\mathbf{w}} h^l(\mathbf{v}) = \langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}}\mathcal{M}(\mathbf{v})\rangle^\top \langle \mathbf{A}^{\otimes l}, \mathcal{M}(\mathbf{w}) - \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})\rangle \tag{20}$$

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

Therefore, since the first-order optimality condition for (5) is that $D_{\mathbf{w}} h^l(\mathbf{v}) = 0$ $\forall \mathbf{v} \in \mathbb{R}^{nr\circ l}$, it can be equivalently written as

$$\langle \langle \mathbf{A}^{\otimes l}, \mathbf{P}(\mathbf{w})\rangle_{2*[l]}, \langle \mathbf{A}^{\otimes l}, \mathcal{M}(\mathbf{w}) - \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})\rangle\rangle_{1,3,\ldots,2l-1} = 0, \tag{21}$$

and left-hand side of the above equation yields (17a) after rearrangements.

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

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

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

$$\nabla_{\mathbf{w}} h^l(\mathbf{P}(\mathbf{w})) = \langle \nabla h^l(\mathbf{X}), \mathbf{P}^{\otimes l}\rangle_{1,2,4,5,\ldots,3l-1,3l} \tag{22}$$

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

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

Combined with (22), we conclude that

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

which yields (17b) directly. □

Now, we turn to showcasing the relationship between the FOPs of (5) and those of (2), which also have a one-to-one correspondence in the symmetric rank-1 regime. This is the reason why it is necessary to introduce (5) despite the extra complication, as rank-1 components tensors in $\mathbb{R}^{[n\times r]\circ l}$ are not lifted versions of $X \in \mathbb{R}^{n\times r}$.

Theorem 3. *For the lifted formulation* (5)*, the first-order condition* $\nabla h^l(\hat{\mathbf{w}}) = 0$ *holds for a symmetric rank-1 tensor* $\hat{\mathbf{w}}$ *if and only if*

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

where $\hat{X} \in \mathbb{R}^{n\times r}$ *is an FOP of* (2)*.*

Proof of Theorem 3. When $\hat{\mathbf{w}} = \mathrm{vec}(\hat{X})^{\otimes l}$, Lemma A.4 and (17a) together imply that

$$\langle \nabla f^l(\langle \hat{X}^{\otimes l}, \hat{X}^{\otimes l}\rangle_{2*[l]}), \hat{X}^{\otimes l}\rangle_{2*[l]} = (\nabla f(\hat{X}\hat{X}^\top)\hat{X})^{\otimes l} = 0 \tag{23}$$

which is equivalent to

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

which is exactly (13). □

Theorem 3 establishes a robust connection between the first-order critical points of the lifted formulation and those of the unlifted formulation. This implies that when first-order methods approach a critical point in (5), valuable information about an FOP of (2) can also be readily extracted. However, the primary challenge in optimizing (2) stems from spurious solutions, which cannot be escaped by first or even second-order algorithms. Consequently, it becomes crucial to examine whether the Hessians of the FOPs of (5), especially those that correspond to the spurious solutions of (2), exhibit any unique properties. As it turns out, the non-global FOPs of (5) display some highly favorable characteristics: they no longer constitute second-order critical points of (5) and are transformed into strict saddles when the parametrization level l is sufficiently large.

To motivate our analysis of conversion from spurious solutions to strict saddle points, we first offer a closer analysis to the SOPs of the unlifted problem (2), which also serves as the key intuition into our main results in this section.

The main observation is that, for a spurious SOP $\hat{X}$ and any ground truth Z with $\hat{X}\hat{X}^\top \neq ZZ^\top$, although they all obey conditions (13) and (14), they still have intrinsic differences that can be amplified via over-parametrization. To illustrate this phenomenon in more detail, we will introduce the following Lemma:

Lemma 8. *For an arbitrary FOP* $\hat{X} \in \mathbb{R}^{n\times r}$ *of* (2) *satisfying the* (α_s, r)*-RSC property, the following inequality holds:*

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

Proof for Lemma 8. According to [17], $\nabla f(M)$ can be assumed to be symmetric without loss of generality. Hence, one can select $u \in \mathbb{R}^n$ such that $u^\top \nabla f(\hat{x}\hat{x}^\top)u = \lambda_{\min}(\nabla f(\hat{x}\hat{x}^\top))$. Then via the definition of RSC we have

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

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

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

according to (13) and since $f(\hat{X}\hat{X}^\top) - f(M^*) \geq 0$, one can write that

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

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

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

which implies that

$$\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \leq -\alpha_s \frac{\|\hat{X}\hat{X}^\top - M^*\|_F^2}{2\,\mathrm{tr}(M^*)} \leq 0 \tag{25}$$

This completes the proof. □

Now let us recall (14), which can be stated equivalently as

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

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

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

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

$$\alpha_s \frac{\|\hat{X}\hat{X}^\top - M^*\|_F^2}{2\,\mathrm{tr}(M^*)} \geq 4L_s\|\hat{X}U^\top\|_F^2 \quad \forall U \tag{26}$$

which means that the L_s and α_s parameters should be benign, and this essentially constitutes the main proof strategy in the existing literature showing in-existence of spurious solutions under benign RIP or RSS/RSC conditions [14, 17, 46, 15, 45].

Therefore, it is natural to ask, in the case when L_s and α_s do not satisfy (26), whether one can systematically over-parametrize the problem so that the LHS of (26) eventually becomes bigger than the RHS. We know that if we just raise both the RHS and LHS to arbitrary powers, the sign of the inequality will not flip. Therefore, the key insight is that if we keep the constant 4 unchanged, and lift the other terms to arbitrary powers, we can eventually satisfy (14). In general terms, we take the following steps in order to establish a strong result regarding the conversion of spurious solutions to strict saddle points:

1. Proving that $\langle \nabla f^l(\langle \mathbf{P}(\hat{\mathbf{w}}), \mathbf{P}(\hat{\mathbf{w}})\rangle, \Delta \otimes \Delta\rangle \geq |\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top))|^l$ for some appropriately chosen point $\Delta \in \mathbb{R}^{nrol}$.
2. Proving that $\|\langle \mathbf{A}^{\otimes l}, \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\Delta)\rangle_{2*[l]} + \langle \mathbf{P}(\Delta), \mathbf{P}(\mathbf{w})\rangle_{2*[l]}\rangle\|_F^2 \leq 4L_s\|\hat{X}U^\top\|_F^{2l}$ for some appropriately chosen points $\Delta \in \mathbb{R}^{nrol}$ and $U \in \mathbb{R}^{n\times r}$
3. Finding the smallest l that converts the spurious solution to strict saddle point, under mild technical conditions.

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

Theorem 4. *Consider an SOP $\hat{X} \in \mathbb{R}^{n\times r}$ of* (2) *of general rank $r < n$, such that $\hat{X}\hat{X}^\top \neq M^*$, and assume that* (2) *satisfies the RSC and RSS conditions. Then $\hat{\mathbf{w}} = \mathrm{vec}(\hat{X})^{\otimes l}$ is a strict saddle of* (5) *with a rank-1 symmetric escape direction if $\hat{X}$ satisfies the inequality*

$$\|M^* - \hat{X}\hat{X}^\top\|_F^2 \geq \frac{L_s}{\alpha_s}\lambda_r(\hat{X}\hat{X}^\top)\,\mathrm{tr}(M^*) \tag{27}$$

and l is odd and is large enough so that

$$l > \frac{1}{1 - \log_2(2\beta)} \tag{28}$$

where β is defined as

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

Here, L_s and α_s are the respective RSS and RSC constants of (2)*.*

Proof of Theorem 4. By Lemma 8, we select $u \in \mathbb{R}^n$ such that $u^\top \nabla f(\hat{X}\hat{X}^\top)u = \lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top))$ with $\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)) \leq 0$.

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

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

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

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

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

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

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

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

since $\hat{X}^\top u = 0$ according to the first-order condition (13). Therefore,

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

Now, if we choose $\Delta = \mathrm{vec}(U)^{\otimes l}$ for the aforementioned $U \in \mathbb{R}^{n\times r}$, the LHS of (17b) can be expressed as:

$$\begin{aligned}\text{LHS} &= 2(\langle \mathbf{A}, \hat{X}\hat{X}^\top\rangle_{2,3}^\top\langle \mathbf{A}, uu^\top\rangle_{2,3})^l - 2(\langle \mathbf{A}, M^*\rangle_{2,3}^\top\langle \mathbf{A}, uu^\top\rangle_{2,3})^l + 4(\|\langle \mathbf{A}, \hat{X}U^\top\rangle_{2,3}\|_2^2)^l\\ &\le 2(\lambda_{\min}(\nabla f(\hat{X}\hat{X}^\top)))^l + 4C_2^l\\ &= 2C_1^l + 4C_2^l\end{aligned} \tag{29}$$

where the inequality follows from:

$$a^n - b^n \le (a-b)^n, \quad \forall b \ge a \ge 0$$

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

$$\text{LHS of (17b)} \le \underbrace{-2G^l}_{\text{Part 1}} + \underbrace{\frac{2}{2^{l-1}}L_s^l\lambda_r(\hat{X}\hat{X}^\top)^l}_{\text{Part 2}}$$

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

$$G^l > (1/2^{l-1})L_s^l\lambda_r(\hat{X}\hat{X}^\top)^l \tag{30}$$

To derive a sufficient condition for (30), we first need a lower bound on G, and Lemma (8) conveniently provides this bound, giving that

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

Therefore, if

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

we can conclude that (30) holds, which implies that the LHS of (17b) is negative, directly proving that $\hat{X}^{\otimes l}$ is not an SOP anymore. Elementary manipulations of the above equation give that a sufficient condition is

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

We now consider (27), which means that

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

Subsequently, define a constant γ such that

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

Then, according to Lemma 5 and (31), we can conclude that $\gamma \ge 1$. Moreover, (33) also means that $\gamma < 2$. 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 (17b) is negative, proving that $\mathrm{vec}(\hat{X})^{\otimes l}$ is a saddle point with the escape direction $\mathrm{vec}(U)^{\otimes l}$, proving the claim.

Next, we aim to study how large l needs to be in order for (34) to hold. Again, we know that

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

and that $\beta \leq 1$ due to assumption (27). Therefore, for (34) to hold true, it is enough to have

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

□

B.1 Other Considerations of Lifted Landscape

In the previous sections, we have shown that by lifting the optimization problem (2) into tensor spaces, we could convert spurious local solutions into strict saddle points. However, it is also important that we could distinguish the true ground truth solutions $Z \in \mathbb{R}^{n\times r}$ with $ZZ^\top . = M^*$ from the spurious ones. This requires that the true solutions Z will remain SOPs after lifting, which we indeed prove in the following theorem:

Theorem 5. *Assume that $Z \in \mathbb{R}^{n\times r}$ is a ground truth solution of* (2) *such that $ZZ^\top = M^*$. Then $\mathrm{vec}(Z)^{\otimes l}$ remains an SOP of* (5) *regardless of the parametrization level l, and without the need for* (2) *to satisfy the RSC or RSS conditions.*

Proof of Theorem 5. Let us start with the first-order optimality condition. Consider the linear map in the proof of Lemma 7 $\mathcal{M} : \mathbb{R}^{nr\circ l} \mapsto \mathbb{R}^{n\circ 2l}$

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

Again, it is apparent that

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

Therefore, at the point $\mathbf{M} = \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})$, we know that $\nabla f^l(\mathcal{M}(\mathrm{vec}(Z)^{\otimes l})) = 0$. Consequently, the LHS of (17a) is equal to zero since it is a product between $\nabla f^l(\mathcal{M}(\mathrm{vec}(Z)^{\otimes l}))$ and $\mathbf{P}(\mathrm{vec}(Z)^{\otimes l})$.

Next, we turn to the second-order optimality condition. Again, recall from the proof of Lemma 7 that

$$\text{LHS of (17b)} = \underbrace{2\langle \nabla f^l(\mathcal{M}(\mathbf{w})), \mathcal{M}(\boldsymbol{\Delta}) \rangle}_{\text{Part 1}} + \underbrace{\|\langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}} \mathcal{M}(\boldsymbol{\Delta}) \rangle\|_F^2}_{\text{Part 2}}$$

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

$$\text{LHS of (17b)} = \|\langle \mathbf{A}^{\otimes l}, D_{\mathbf{w}} \mathcal{M}(\boldsymbol{\Delta}) \rangle\|_F^2 \geq 0, \qquad \forall \boldsymbol{\Delta}$$

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

Next, it is important to analyze the main results obtained in Theorem 4 under the lens of some other existing characterizations of the loss landscape of (2). According to Theorem 4, over-parametrization or lifting proves to be highly beneficial when dealing with spurious solutions, represented by $\hat{X}$, that significantly deviate from the actual ground truth. The theorem implies that as the distance between $\hat{X}$ and the ground truth increases, a smaller value of l is necessary for $\mathrm{vec}(\hat{X})^{\otimes l}$ to evolve into a saddle point, as alluded to in (28). This concept is consistent with previous studies, which maintain that the area surrounding M^* exhibits a favorable optimization landscape, characterized by an absence of deceptive local solutions in a specified zone around M^*. A commonly cited illustration of this assertion is provided below.

Theorem 6 (Theorem 3 [47]). *If $\hat{X}$ is an SOP of* (2) *and*

$$\|\hat{X}\hat{X}^\top - M^*\|_F \leq \frac{4L_s\alpha_s}{(L_s+\alpha_s)^2}\lambda_r(M^*), \tag{35}$$

then

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

This means that any spurious solution of (2) is reasonably far away from the ground truth solution M^*. Coupled with the fact that lifting the problem into higher-dimensional tensor spaces can convert spurious solutions far away from M^* into strict saddles points, we can ascertain that by setting the RHS of (35) to be greater or equal to the RHS of (27), all spurious solutions will be converted into strict saddle points via lifting. We make this key observation concrete in the following theorem.

Theorem 7. *Assume that $\hat{X} \in \mathbb{R}^{n\times r}$ is a spurious solution of* (2)*, and that* (2) *satisfies the RSC and RSS assumptions with the α_s and L_s constants, respectively. Then $\mathrm{vec}(\hat{X})^{\otimes l}$ is a strict saddle point of* (5) *for an odd l satisfying* (28) *if*

$$\|M^*\|_F \leq \frac{1}{\tau\sqrt{r}}\frac{2\sqrt{2}\alpha_s^{5/2}}{(L_s+\alpha_s)^2\sqrt{L_s}} \tag{36}$$

where τ is the condition number of M^*.

Proof of Theorem 7. By using Lemma 6, we know that

$$\text{RHS of (27)} \leq \sqrt{\frac{2L_s^3}{r\alpha_s^3}}\|M^*\|_F \,\mathrm{tr}(M^*)$$

Hence, it is enough to make the RHS of the above inequality to be less than that of (35), meaning that

$$\sqrt{\frac{2L_s^3}{r\alpha_s^3}}\|M^*\|_F \,\mathrm{tr}(M^*) \leq \frac{4L_s\alpha_s}{(L_s+\alpha_s)^2}\lambda_r(M^*) \implies \|M^*\|_F\frac{\mathrm{tr}(M^*)}{\lambda_r(M^*)} \leq \frac{2\sqrt{2r}\alpha_s^{5/2}}{(L_s+\alpha_s)^2\sqrt{L_s}}$$

Then, acknowledging that $\mathrm{tr}(M^*) \leq r\tau\lambda_r(M^*)$ completes the proof. □

C Additional Details for Implicit Bias of GD in Tensor Space

C.1 More Tensor Algebra

Definition 9. Given a cubic tensor $\mathbf{w} \in \mathbb{R}^{n \circ l}$, its spectral norm $\|\cdot\|_S$ and nuclear norm $\|\cdot\|_*$ are defined respectively as

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

From the definition, it also follows that

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

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

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

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

Lemma 9 (Theorem 2.1, 3.2 [38])**.** *For tensors* $\mathbf{w}$ *and* $\mathbf{v}$ *of appropriate dimensions (if doing inner product, the dimensions along which the multiplication is performed must have matching size), we have*

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

Moreover, they have a dual norm relationship:

Lemma 10 (Lemma 21 [48])**.** *The spectral norm* $\|\cdot\|_S$ *is the dual norm to the nuclear norm* $\|\cdot\|_*$*, namely given an arbitrary tensor* $\mathbf{w}$*, we have that*

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

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

Next, we introduce the notion of eigenvalues for tensors. There are many related definitions, like outlined in [39]. However, we introduce a novel variational characterization of eigenvalues that resembles the Courant-Fisher minimax definition for eigenvalues of matrices. Note this is a new definition that is fist introduced in this paper, and may be of independent interest outside of the current scope.

Definition (Definition 4, Variational Eigenvalue of Tensors)**.** For a given tensor $\mathbf{w} \in \mathbb{R}^{n \circ l}$, we define its k^{th} variational eigenvalue (v-Eigenvalue) $\lambda_k^v(\mathbf{w})$ as

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

where S is a subspace of $\mathbb{R}^{n \circ l}$ that is spanned by a set of orthogonal, symmetric, rank-1 tensors. Its dimension denotes the number of orthogonal tensors that span this space.

It is apparent from the definition that $\|\mathbf{w}\|_S = \lambda_1^v(\mathbf{w})$. Note that our definition of v-Eigenvalues of tensors can only define n eigenvalues at most, which is not the maximum amount of H- or Z-Eigenvalues a tensor can have [39], and it is well known that even with symmetric tensors, its rank can go well beyond n [49]. We also note that this definition exactly coincides with the definition of Hermitian tensor eigenvalues (introduced here [50]) when constrained to Hermitian tensors [51]. We also conjecture that this definition coincides with the top-n Z-Eigenvalues for even-order symmetric real tensors [39], but it is an open question for now.

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

Proposition 3. *For an integer k in $[1,\ldots,n]$, the k^{th} variational eigenvalue (v-Eigenvalue) $\lambda_k^v(\mathbf{w})$ of a tensor $\mathbf{w}$ satisfies:*

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

Proof of Proposition 3. We prove the proposition by contradiction. Assume that the two formulations claimed to be identical in Proposition 3 are not the same. We further assume that S is spanned by symmetric, rank-1 tensors $\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}$, and that T is spanned by symmetric, rank-1 tensors $\{\mathbf{u}_{-(n-k+1)},\ldots,\mathbf{u}_{-1}\}$, meaning that

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

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

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

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

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

It follows that

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

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

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

and also

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

as otherwise the *outer* maximization formulation affecting the choice of u_k will make $\xi_2 = 0$, contradicting our claim. By definition we have $\text{span}\{\mathbf{u}_1,\ldots,\mathbf{u}_k\} \bigcap \text{span}\{\mathbf{u}_{-(n-k)},\ldots,\mathbf{u}_{-1}\} = \{\emptyset\}$.

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

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

Lemma 11 (Tensor Weyl's)**.** *Consider two tensors $\mathbf{w}$ and $\mathbf{v}$ of the same dimension. It holds that*

$$\lambda_k^v(\mathbf{w}) + \lambda_1^v(\mathbf{v}) \geq \lambda_k^v(\mathbf{w}+\mathbf{v}) \geq \lambda_k^v(\mathbf{w}) - \lambda_1^v(\mathbf{v}) \tag{37}$$

The proof of Lemma 11 is highly similar to that of Theorem 2 in [51], only substituting for our new definition of v-Eigenvalues, thus omitted for simplicity.

C.2 Main Results and Their Proofs

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

We start with the proof to Lemma 1.

Proof of Lemma 1. We proceed with the proof by induction. First, assume that $\mathbf{w}_0 = x_0^{\otimes l}$ for some $x_0 \in \mathbb{R}^{nr}$. One can write

$$\nabla h^l(\mathbf{w}_0) = \langle\langle (I_r \oslash_{1,2} \mathbf{A})^{\otimes l}, \mathbf{w}_0\rangle_{2*[l]}, \langle \mathbf{A}^{\otimes l}, \mathcal{M}(\mathbf{w}_0) - \mathcal{M}(\mathrm{vec}(Z)^{\otimes l})\rangle\rangle_{1,3,\ldots,2l-1} \tag{38}$$

where $\mathcal{M}(\cdot)$ is defined per proof of Lemma 7. The difference between this formulation and (21) is that we have replaced $\langle \mathbf{A}^{\otimes l}, \mathbf{P}(\mathbf{w}_0)\rangle_{2*[l]}$ with $\langle (I_r \oslash_{1,2} \mathbf{A})^{\otimes l}, \mathbf{w}_0\rangle_{2*[l]}$, which are equivalent, just with the second tensor having the dimensions $nr, m, \ldots, nr, m$ so that $\nabla h^l(\mathbf{w}_0)$ has the dimensions $nr, \ldots, nr$. Note that $\oslash$ denotes the usual kronecker product, which can be thought of a reshaped version of tensor outer product. $\oslash_{1,2}$ denotes the kronecker product only happening with respect to the first 2 dimensions of $\mathbf{A}$. From now on, we denote $\mathbf{A}_r \coloneqq I_r \oslash_{1,2} \mathbf{A}$.

Now, according to the above formulation and Lemma A.4, we have

$$\begin{aligned} \nabla h^l(\mathbf{w}_0) &= \left(\langle \mathbf{A}_r, \langle \mathbf{A}, \mathrm{mat}(x_0)\,\mathrm{mat}(x_0)^\top - M^*\rangle\rangle_{3,6,\ldots,3l}\, x_0\right)^{\otimes l} \\ &\coloneqq \left(\langle \mathbf{A}_r^* \mathbf{A}, \mathrm{mat}(x_0)\,\mathrm{mat}(x_0)^\top - M^*\rangle\, x_0\right)^{\otimes l} \end{aligned} \tag{39}$$

where

$$(\mathbf{A}_r^l)^* \mathbf{A}^l \coloneqq \langle (\mathbf{A}_r)^{\otimes l}, \mathbf{A}^{\otimes l}\rangle_{3,6,\ldots,3l} \in \mathbb{R}^{[nr\times nr\times n\times n]\circ l} \tag{40}$$

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

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

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

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

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

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

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

Lemma 12. *The GD trajectory of* (5) $\{\mathbf{w}_t\}_{t=0}^\infty$ *admits the following breakdown for an arbitrary* t*:*

$$\mathbf{w}_{t+1} = \langle \mathbf{Z}_t, \mathbf{w}_0\rangle - \mathbf{E}_t \coloneqq \tilde{\mathbf{w}}_t - \mathbf{E}_t \tag{41}$$

where

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

and where $(\mathbf{A}_r^l)^* \mathbf{A}^l \coloneqq \langle (\mathbf{A}_r)^{\otimes l}, \mathbf{A}^{\otimes l}\rangle_{3,6,\ldots,3l} \in \mathbb{R}^{[nr\times nr\times n\times n]\circ l}$.

Proof of Lemma 12. For this proof, we will proceed by induction. For $t = 1$, we have that

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

Then, we move on to the induction step, while first assuming that it holds for some t. One can write

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

□

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

Lemma 13. *Given a tensor $\mathbf{E}_t$ defined in Lemma 12, assume that $\mathbf{w}_0 = \epsilon x_0^{\otimes l}$, where $\epsilon \in \mathbb{R}$ is the initialization scale. For every $t \leq t_s$,*

$$\|\mathbf{E}_t\|_S \leq \frac{8}{r_U^l \sigma_1(U)^l}\epsilon^3(nL_s)^{l/2}(1 + \tilde{\eta}\sigma_1(U)^l)^{3t}\|x_0^{\otimes l}\|_*^3 \tag{42}$$

with

$$t_s = \lfloor \frac{\ln\left(\frac{\sigma_1^l(U) r_U^l}{8r^l L_s^{l/2}\|x_0^{\otimes l}\|_*^3}\frac{|x_0^\top v_1|^l}{n^{l/2}}\right) - 2\ln(\epsilon)}{2\ln(1 + \tilde{\eta}\sigma_1^l(U))} \rfloor \tag{43}$$

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

Proof of Lemma 13. From Lemma 9 and the definition in Lemma 12, it is apparent that

$$\|\mathbf{E}_t\|_S \leq \sum_{i=1}^{t}\|(\mathcal{I} + \eta\langle(\mathbf{A}_r^l)^*\mathbf{A}^l, (M^*)^{\otimes l}\rangle)^{t-i}\|_S\|\hat{\mathbf{E}}_i\|_* \tag{44}$$

We proceed to derive upper bounds on the norm terms separately, and then combine them together later. We first deal with $\|\hat{\mathbf{E}}_i\|_*$. By Lemma 9, we have that

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

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

$$\mathbf{w}_{i-1} = \sum_{j=1}^{r_{i-1}}\lambda_j(x_j^{i-1})^{\otimes l}, \quad x_j^{i-1} \in \mathbb{R}^{nr},\ \|x_j^{i-1}\|_2 = 1 \tag{45}$$

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

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

leading to

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

For given indices j_1, j_2 index, it follows from Lemma A.4 that

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

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

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

Hence, the eigenvalues of the LHS are just r copies of that of the RHS [44]. This further implies

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

where the second last inequality follows from the RSS property, and the last equality follows from (45). Next, we apply Lemma 9 again with

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

which leads to

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

This directly gives

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

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

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

where $\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, (M^*)^{\otimes l} \rangle \in \mathbb{R}^{nr \circ 2l}$, and $(\cdot)^k$ just denotes repeated multiplications along the even dimensions of the tensor, as explained in the disclaimer. To upper-bound the spectral norm of $(\mathcal{I} + \eta \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, (M^*)^{\otimes l} \rangle)^{t-i}$, it is necessary to upper-bound the spectral norm of $(\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, (M^*)^{\otimes l} \rangle)^k$. To do so, we use Lemma 10 to reformulate

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

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

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

with $\sum_{j_v} |\lambda_{j_v}| = \|\mathbf{v}^*\|_* \leq 1$. Note that this decomposition is due to the fact that $\mathbf{v}$ is not necessarily symmetric. Again, by Lemma A.4,

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

directly leading to

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

Since

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

this means that

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

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

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

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

$$\|\tilde{\mathbf{w}}_t - \mathbf{w}_t\|_* \leq \|\tilde{\mathbf{w}}_t\|_*, \quad \forall t \leq t_s \tag{46}$$

where $\tilde{\mathbf{w}}_t$ is defined in (41). We will later justify the existence of t_s and derive a lower bound. If the above inequality holds true, we also have

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

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

$$\tilde{\mathbf{w}}_t = \sum_{k=0}^{t} \binom{t}{k} \eta^k \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, (M^*)^{\otimes l} \rangle^k \mathbf{w}_0 \tag{47}$$

Therefore, it follows from Lemma 9 that,

$$\|\tilde{\mathbf{w}}_{i-1}\|_* \leq \left(\sum_{k=0}^{i-1} \binom{i-1}{k} \eta^k \|\langle (\mathbf{A}_r^l)^* \mathbf{A}^l, (M^*)^{\otimes l} \rangle^k\|_* \right) \|\mathbf{w}_0\|_* \tag{48}$$

for all $i \leq t$. With the repeated application of Lemma 9, we have

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

Therefore, substituting back into (48) gives

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

Next, plugging the above preparatory results into (44), we have that

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

proving the original claim of this lemma (42). Now, we give a lower bound on t_s. By recalling the breakdown (47), we have

$$
\begin{aligned}
\|\tilde{\mathbf{w}}_t\|_* \geq \|\tilde{\mathbf{w}}_t\|_S &\geq \langle\tilde{\mathbf{w}}_t, v_1^{\otimes l}\rangle \\
&= \epsilon\sum_{k=0}^{t}\binom{t}{k}\eta^k\left[|v_1^\top\langle\mathbf{A}_r^*\mathbf{A}, M^*\rangle^k x_0|\right]^l \\
&= \epsilon\sum_{k=0}^{t}\binom{t}{k}\eta^k\left[|v_1^\top U^k x_0|\right]^l \\
&= \epsilon\sum_{k=0}^{t}\binom{t}{k}\eta^k(|\sigma_1^k(U)v_1^\top x_0|)^l = \epsilon|v_1^\top x_0|^l(1+\eta\sigma_1^l(U))^t
\end{aligned} \tag{49}
$$

with v_1 being the first singular vector of $I_r \oslash U$. Since the sensing matrices are assumed to be symmetric, U is also symmetric, hence the singular vectors of U^k coincide with those of U. By (42), we also know

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

Therefore, for (46) to hold true, we need the RHS of the above equation to be smaller than 1, meaning that

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

This further implies that for (46) to hold, t should satisfy

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

which after rearrangement gives (43). □

Now, we present the proof of Lemma 2.

Proof of Lemma 2. Using the tensor Weyl's inequality (Lemma 11), we have that

$$\lambda_2^v(\mathbf{w}_t) \leq \lambda_2^v(\tilde{\mathbf{w}}_t) + \|\mathbf{E}_t\|_S \tag{50}$$
$$\lambda_1^v(\mathbf{w}_t) \geq \lambda_1^v(\tilde{\mathbf{w}}_t) - \|\mathbf{E}_t\|_S \tag{51}$$

The only remaining part of the proof is the characterization of $\lambda_1^v(\tilde{\mathbf{w}}_t)$ and $\lambda_2^v(\tilde{\mathbf{w}}_t)$. The first term is easy because we already have the characterization from the proof of Lemma 13, with (49) giving rise to

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

Also, by the definition of v-eigenvalues and (47), we have that

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

where v_2 is the singular vector associated with $\sigma_2^k(U)$ $\forall k \in [t]$. Finally, combining the above equations yields (6) after rearrangements. □

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

Lemma 14. *Let $x_0 = v_1 + g \in \mathbb{R}^{nr}$, where g is a vector with each entry being i.i.d sampled from Gaussian distribution $\mathcal{N}(0, \rho)$. For some universal constant C, the follwoing inequalities hold:*

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

Proof of Lemma 14. We know that

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

Theorem 2.6.3 of [52] (general Hoeffding's) gives that with probability at least $1 - 2\exp(-t^2/\rho^2)$,

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

which leads to the first concentration bound after substituting $t = \mathcal{O}(\sqrt{\rho})$ with some constant c_1. Then, Theorem 3.1.1 in [52] gives

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

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

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

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

Then, we prove our main theorem of this section.

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

$$\lambda_2^v(\tilde{\mathbf{w}}_t) \leq \frac{\zeta}{2}\lambda_1^v(\tilde{\mathbf{w}}_t), \tag{52}$$

$$\|\mathbf{E}_t\|_s \leq \frac{\zeta}{2}\lambda_1^v(\tilde{\mathbf{w}}_t) \tag{53}$$

By recalling Lemma 2, a sufficient condition for (52) is that

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

implying that

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

which after rearrangements gives $t \geq t(\zeta, l)$, as defined in (9). Then, we obtain a sufficient condition for (53), which by Lemma 13 is

$$\frac{8r^l}{r_U^l\sigma_1(U)^l}\epsilon^3(nL_s)^{l/2}(1+\tilde{\eta}\sigma_1(U)^l)^{3t}\|x_0^{\otimes l}\|_*^3 \leq \frac{2}{\zeta}\epsilon|v_1^\top x_0|^l(1+\eta\sigma_1^l(U))^t \tag{54}$$

contingent on the fact that $t \leq t_s$. Therefore, before going further, we need to verify that $t(\zeta, l) \leq t_s$ for some small enough ϵ. (43) implies that a sufficient condition is

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

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

$$\frac{\ln(1+\tilde{\eta}\sigma_1^l(U))}{\ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1}} \leq \frac{r_U^l(1+\eta\sigma_1^l(U))}{1-(\sigma_2(U)/\sigma_1(U))^l} \coloneqq \Xi \tag{55}$$

Hence,

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

and after rearrangement gives

$$\epsilon^2 \leq \frac{\sigma_1^l(U)r_U^l}{8(r^2nL_s)^{l/2}}\frac{|x_0^\top v_1|^l}{\|x_0^{\otimes l}\|_*^3}\left(\frac{2\|x_0\|_2^l}{\zeta|v_1^\top x_0|^l}\right)^{-\Xi} \tag{56}$$

Notice that all of the above terms are independent of ϵ, and are positive. Therefore, a small enough ϵ exists. Also notice that a smaller step-size η will yield a loser bound on ϵ through the dependence of Ξ. Now, consider (54) again. Since T is finite, a sufficient condition for (54) is

$$\epsilon^2 \leq \zeta\frac{r_U^l\sigma_1(U)^l}{16(r^2nL_s)^{l/2}}\frac{|v_1^\top x_0|^l}{\|x_0^{\otimes l}\|_*^3}\left(\frac{1+\eta\sigma_1^l(U)}{(1+\tilde{\eta}\sigma_1(U)^l)^3}\right)^T \tag{57}$$

which can again be achieved by setting a small enough ϵ, since all other terms are positive and not dependent on it. In summary, if we choose a small constant ϵ satisfying both (56) and (57), and if $t_s \geq t_T$ (which again can be achieved via a sufficiently small ϵ), it is already sufficient for both (52) and (53) to hold, thereby giving:

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

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

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

As a result,

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

which proves (8). □

Theorem 1 can also be improved via Lemma 14 as stated below.

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

$$\frac{\lambda_2^v(\mathbf{w}_t)}{\lambda_1^v(\mathbf{w}_t)} \leq \kappa \quad \textit{for } t \asymp \ln\left(\frac{1}{\kappa}\right)\ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1} \tag{58}$$

provided that

$$\epsilon \asymp \sqrt{\kappa/2}\frac{(\sigma_1(U)r_U)^{l/2}}{4(r^2nL_s)^{l/4}}(\frac{\kappa}{4})^{3\Xi/2}, \quad \textit{where } \Xi := \frac{r_U^l(1+\eta\sigma_1^l(U))}{1-(\sigma_2(U)/\sigma_1(U))^l} \tag{59}$$

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

Proof of Corollary 1. The proof is similar to that of Theorem 1 (note $\zeta = \kappa/2$), and therefore we only highlight the difference. We know that (23) holds true if

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

It results from Lemma 14 that for our choice of initialization, we have that

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

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

$$t \asymp \ln\left(\frac{2}{\zeta}\right)\ln\left(\frac{1+\eta\sigma_1^l(U)}{1+\eta\sigma_2^l(U)}\right)^{-1} := t_*, \tag{60}$$

$$\epsilon \asymp \frac{(\sigma_1(U)r_U)^{l/2}}{2\sqrt{2}(r^2nL_s)^{l/4}}\left(\frac{2}{\zeta}\right)^{-\Xi/2} \tag{61}$$

(23) will hold with high probability. Next, in order for (53) to hold for $t \asymp t_*$, we know that

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

Via the same order of magnitude argument, we know that the following condition is sufficient for (53):

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

Now,

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

where the second equality follows from the substitution of t_*, and the last inequality follows from (55). As a result,

$$\epsilon \asymp \sqrt{\zeta}\frac{(\sigma_1(U)r_U)^{l/2}}{4(r^2nL_s)^{l/4}}(\frac{\zeta}{2})^{3\Xi/2} \tag{62}$$

Therefore, taking the minimum of (61) and (62), we know that

$$\epsilon \asymp \sqrt{\zeta}\frac{(\sigma_1(U)r_U)^{l/2}}{(r^2nL_s)^{l/4}}(\frac{\zeta}{2})^{3\Xi/2} \tag{63}$$

is sufficient for (52) and (53), leading to (58) via the same steps in the proof of Theorem 1. □

D Additional Details for Properties of Approximate Rank-1 Tensors

We start with the proof of Proposition 1.

Proof of Proposition 1. Given a symmetric tensor $\mathbf{w}$, it can be decomposed as

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

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

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

and it is apparent that the second term is orthogonal to $w_s^{\otimes l}$ via Lemma A.4. Therefore, we just organize all components $w_s^{\otimes l}$ together and all orthogonal components together. By definition, the parallel component has the magnitude $\lambda_1^v(\mathbf{w})$. Also, by the definition of v-eigenvalues, $\|\mathbf{w}^\dagger\|_S \leq \lambda_2^v(\mathbf{w}_t)$ since otherwise the dominant direction of $\mathbf{w}^\dagger$ will just become the second eigenvector of $\mathbf{w}$. □

We now provide the proof of Proposition 2.

Proof of Proposition 2. According to (17a), the gradient of (5) with respect to $\mathbf{w}$ can be expressed as

$$\nabla h^l(\mathbf{w}) = \langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \langle \mathbf{P}(\mathbf{w}), \mathbf{P}(\mathbf{w}) \rangle_{2*[l]} - (M^*)^{\otimes l} \rangle, \mathbf{w} \rangle_{2*[l]} \tag{64}$$

where $(\mathbf{A}_r^l)^* \mathbf{A}^l$ is defined in (40). In light of (10), one can write

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

where $\mathbf{w}_\sigma = \lambda_1^v(\mathbf{w}) \hat{w}^{\otimes l}$. Note that we have dropped the subscripts from the second line and henceforth for sake of simplicity. By using this logic, (64) can be written as

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

where

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

The first term can be analyzed as

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

and by Lemma 9, we have that

$$\begin{aligned}
\|\langle \langle (\mathbf{A}_r^l)^* \mathbf{A}^l, \mathbf{a}_1 \rangle, \mathbf{w}^\dagger \rangle\|_S &\leq \|\langle \langle \mathbf{P}^{\otimes l}, \mathbf{P}^{\otimes l} \rangle, \mathbf{w}_\sigma \otimes \mathbf{w}_\sigma \otimes \mathbf{w}^\dagger \rangle\|_S \|(\mathbf{A}_r^l)^* \mathbf{A}^l\|_* \\
&= \|\langle \mathbf{P}(\mathbf{w}_\sigma), \mathbf{P}(\mathbf{w}_\sigma) \rangle \otimes \mathbf{w}^\dagger\|_S \|(\mathbf{A}_r^l)^* \mathbf{A}^l\|_* \\
&\leq \lambda_1^v(\mathbf{w})^2 \|\mathbf{w}^\dagger\|_S \|(\mathbf{A}_r^l)^* \mathbf{A}^l\|_* \leq \kappa \lambda_1^v(\mathbf{w})^3 r^l \|\mathbf{A}^* \mathbf{A}\|_*^l
\end{aligned} \tag{65}$$

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

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

Repeating this process leads to

$$\|\mathbf{h}_2\|_S \leq (3\kappa + 3\kappa^2 + \kappa^3 + \kappa\|M^*\|_F^2)\lambda_1^v(\mathbf{w})^3 r^l \|\mathbf{A}^*\mathbf{A}\|_*^l \tag{66}$$

Similarly, $\|\mathbf{h}_1\|_S = \mathcal{O}(\lambda_1^v(\mathbf{w})^3 r^l \|\mathbf{A}^*\mathbf{A}\|_*^l)$. Now, if we assume that $\mathbf{w}$ is an FOP of (5), it means that $\nabla h^l(\mathbf{w}) = 0$, further implying $\|\nabla h^l(\mathbf{w})\|_S = 0$, and by reverse triangle inequality,

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

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

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

which is equivalent to the FOP condition for (2), which is (13), meaning that $\mathrm{mat}(w_s) \in \mathbb{R}^{n\times r}$ is an FOP of (2). Note that we can always scale $\mathbf{A}$ and b together so that $\|\mathbf{A}^*\mathbf{A}\|_*^l$ can be normalized to 1. □

Finally, we prove the main result of this paper.

Proof of Theorem 2. We consider the SOP condition for (5), which is (17b) for some rank-1 tensor Δ. We can express it as

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

Let Δ be defined identically to that in the proof of Theorem 4, meaning that $\Delta = \mathrm{vec}(U)^\otimes := u^{\otimes l}$. By the same logic of (64), we have that

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

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

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

Hence,

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

Now, we turn to $\mathbf{a}_2(\hat{\mathbf{w}})$. Since the sensing matrices are assumed to be symmetric, by (29), we have

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

again following the procedures in (64). Given the decomposition of $\hat{\mathbf{w}}$, we decompose $\mathbf{b}_2$ similarly to $\mathbf{b_1}$:

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

Combining everything together, we have

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

In addition, following the same procedures in (65),

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

Now, since $\mathbf{w}_\sigma$ is a lifted version of FOP for (2) (via Proposition 1),

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

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

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

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

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

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

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

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

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

and this leads to the desirable result. □

E Additional Experiments

In this section, we provide some additional experiments to showcase the algorithmic regularization of GD algorithm in tensor problems like (5).

This section involves the decomposition of tensors along the optimization trajectory using a known algorithm, S-HOPM, as outlined in [23]. The S-HOPM algorithms extract the dominant rank-1 component of a given tensor, so as a first step, we apply this to tensors on the trajectory, and obtain $\mathbf{w}_1$. Subsequently, this component was subtracted from the original tensor, and the extraction procedure was repeated on the resultant tensor $\mathbf{w} - \mathbf{w}_1$ to obtain a new component $\mathbf{w}_2$. This allows us to directly compute $\frac{\|\mathbf{w}_1\|_F}{\|\mathbf{w}_2\|_F}$, in the hope to approximate $\lambda_2^v(\mathbf{w}_t)/\lambda_1^v(\mathbf{w}_t)$ for some given t in the trajectory. Note that this procedure mirrors the definition of the variational eigenvalue of tensors defined in Definition 4. The main source of inaccuracy is that the S-HOPM algorithm may not find the real dominant rank-1 component, as specified in the original paper. Therefore, the metric we show below only serves as an approximation of $\lambda_2^v(\mathbf{w}_t)/\lambda_1^v(\mathbf{w}_t)$.

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

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

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

This ratio provides meaningful insights into the training dynamics, which further substantiates the claims made under Theorem 1.

F Custom Algorithms

Algorithm 1: CustomGD Algorithm

Input: learning_rate, n, r, l, prob_params, loss, g_thres, buffer, beta, gamma, eta_0
Initialize variables: A, b, escape_saddle, buffer_limit, buffer_step
Function `init`(*starting_point, lr*)
    if *lr* $\neq$ *0* **then**
        learning_rate $\leftarrow$ lr `// Update learning rate if specified`
    end
    return $\{'curr\_iter' : 0, 't\_noise' : 0, 'curr\_w' : starting\_point\}$
Function `update`(*gradients, opt_state*)
    curr_iter $\leftarrow$ opt_state['curr_iter'] + 1
    t_noise $\leftarrow$ opt_state['t_noise']
    curr_w $\leftarrow$ opt_state['curr_w']
    if $\|$*gradients*$\|$ $< g\_thres$ *and curr_iter* $>$ *100* **then**
        if *escape_saddle* **then**
            t_noise $\leftarrow$ curr_iter
            w_s $\leftarrow$ find rank 1 component of curr_w using tensor PCA
            direction $\leftarrow$ find the escape direction of w_s `// According to Theorem 2`
            this_eta $\leftarrow$ eta_0
            while *loss(curr_w + this_eta * direction)* $>$ *loss(curr_w) + beta * this_eta * inner_product(gradients, direction)* **do**
                this_eta $\leftarrow$ this_eta * gamma `// Update eta using gamma, backtracking line search`
            end
            updates $\leftarrow$ this_eta * direction
            escape_saddle $\leftarrow$ False
        end
        else
            buffer_step $\leftarrow$ buffer_step + 1
            if *buffer_step* $==$ *buffer_limit* **then**
                escape_saddle $\leftarrow$ True
                buffer_step $\leftarrow$ 0
            end
            updates $\leftarrow$ -learning_rate * gradients
        end
    end
    else
        escape_saddle $\leftarrow$ False
        updates $\leftarrow$ -learning_rate * gradients
    end
    return updates, $\{'curr\_iter'$: curr_iter,$'$ t_noise' : t_noise,$'$ $curr\_w'$: curr_w $+$ updates$\}$

Algorithm 2: Tensor PCA Algorithm

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