Local and Global Convergence of General Burer-Monteiro Tensor Optimizations

Shuang Li,[1] **Qiuwei Li** [2]

[1] Department of Mathematics, University of California, Los Angeles
[2] Alibaba Group (US), Damo Academy
shuangli@math.ucla.edu, liqiuweiss@gmail.com.com

Abstract

Tensor optimization is crucial to massive machine learning and signal processing tasks. In this paper, we consider tensor optimization with a convex and well-conditioned objective function and reformulate it into a nonconvex optimization using the Burer-Monteiro type parameterization. We analyze the local convergence of applying vanilla gradient descent to the factored formulation and establish a local regularity condition under mild assumptions. We also provide a linear convergence analysis of the gradient descent algorithm started in a neighborhood of the true tensor factors. Complementary to the local analysis, this work also characterizes the global geometry of the best rank-one tensor approximation problem and demonstrates that for orthogonally decomposable tensors the problem has no spurious local minima and all saddle points are strict except for the one at zero which is a third-order saddle point.

1 Introduction

Tensors, a multi-dimensional generalization of vectors and matrices, provide natural representations for multi-way datasets and find numerous applications in machine learning and signal processing, including video processing (Liu et al. 2012), hyperspectral imaging (Li et al. 2015b; Sun et al. 2020), collaborative filtering (Hou and Qian 2017), latent graphical model learning (Anandkumar, Ge, and Janzamin 2017), independent component analysis (ICA) (Cardoso 1989), dictionary learning (Barak, Kelner, and Steurer 2015), neural networks compression (Phan et al. 2020; Bai et al. 2021), Gaussian mixture estimation (Sedghi, Janzamin, and Anandkumar 2016), and psychometrics (Smilde, Bro, and Geladi 2005). See (Sidiropoulos et al. 2017) for a review. All these applications involve solving certain optimizations over the space of low-rank tensors:

$$\underset{T}{\text{minimize}}\ f(T) \ \ \text{subject to} \ \ \text{rank}(T) \le r. \tag{1}$$

Here $f(\cdot)$ is a problem dependent objective function with tensor argument and $\text{rank}(\cdot)$ calculates the tensor rank. The rank of matrices is well-understood and has many equivalent definitions, such as the dimension of the range space, or the size of largest non-vanishing minor, or the number of nonzero singular values. The latter is also equal to the smallest number of rank-one factors that the matrix can be written as a sum of. The tensor rank, however, has several non-equivalent variants, among which the Tucker rank (Kolda and Bader 2009) and the Canonical Polyadic (CP) rank (Grasedyck, Kressner, and Tobler 2013) are most well-known. The CP tensor rank is a more direct generalization from the matrix case and is precisely equal to the minimal number of terms in a rank-one tensor decomposition. It is also the preferred notion of rank in applications. Unfortunately, while the Tucker rank can be found by performing the higher-order singular value decomposition (HOSVD) of the tensor, the CP rank is NP-hard to compute (Hillar and Lim 2013). Even though some recent works (Yuan and Zhang 2016; Barak and Moitra 2016; Li et al. 2016; Li and Tang 2017; Li et al. 2015a; Tang and Shah 2015) study the convex relaxation methods based on the tensor nuclear norm, which is also NP-hard to compute (Hillar and Lim 2013). Therefore, this work seeks alternative ways to solve the CP rank-constrained tensor optimizations.

General Burer-Monteiro Tensor Optimizations

Throughout this paper, we focus on third-order, *symmetric* tensors and assume that $f : \mathbb{R}^{n\times n\times n} \to \mathbb{R}$ is a general convex function and has a unique global minimizer $T^\star$ that admits the following (symmetric-)rank-revealing decomposition:

$$T^\star = \sum_{p=1}^{r} c_p^\star \widehat{\boldsymbol{u}}_p \otimes \widehat{\boldsymbol{u}}_p \otimes \widehat{\boldsymbol{u}}_p \in \mathbb{R}^{n\times n\times n}, \tag{2}$$

where $\widehat{\boldsymbol{u}}_p$'s are the normalized tensor factors living on the unit spheres $\mathbb{S}^{n-1}$ and $c_p^\star$'s are the decomposition coefficients. Without loss of generality, we can always assume $c_p^\star > 0$, since otherwise we can absorb its sign into the normalized tensor factors.

Note that the global optimal tensor in (2) can be rewritten as

$$\begin{aligned} T^\star &= \sum_{p=1}^{r} (c_p^{\star 1/3}\widehat{\boldsymbol{u}}_p) \otimes (c_p^{\star 1/3}\widehat{\boldsymbol{u}}_p) \otimes (c_p^{\star 1/3}\widehat{\boldsymbol{u}}_p) \\ &\doteq U^\star \circ U^\star \circ U^\star, \end{aligned} \tag{3}$$

where $U^\star \doteq \left[c_1^{\star 1/3}\widehat{\boldsymbol{u}}_1\ c_2^{\star 1/3}\widehat{\boldsymbol{u}}_2\ \cdots\ c_r^{\star 1/3}\widehat{\boldsymbol{u}}_r\right]$ can be viewed as the "cubic root" of $T^\star$. Noting that the "cubic-root" rep-

resentation (3) has permutation ambiguities, that is, different columnwise permutations of $U^\star$ would generate the same tensor in (3): $[\boldsymbol{u}_{i_1}^\star\ \boldsymbol{u}_{i_2}^\star\ \cdots\ \boldsymbol{u}_{i_r}^\star] \circ [\boldsymbol{u}_{i_1}^\star\ \boldsymbol{u}_{i_2}^\star\ \cdots\ \boldsymbol{u}_{i_r}^\star] \circ [\boldsymbol{u}_{i_1}^\star\ \boldsymbol{u}_{i_2}^\star\ \cdots\ \boldsymbol{u}_{i_r}^\star] = U^\star \circ U^\star \circ U^\star$ for any permutation $(i_1, i_2, \cdots, i_r)$ of the index $(1, 2, \cdots, r)$. This immediately implies that $U^\star$ and its columnwise permutations all give rise to global minimizers of the following reformulation of the optimization (1):

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\ f(U\circ U\circ U). \tag{4}$$

Note that this new factorized formulation has explicitly encoded the rank constraint $\mathrm{rank}(T) \le r$ into the factorization representation $T = U \circ U \circ U$. As a result, the rank-constrained optimization problem (1) on tensor variables reduces to the above unconstrained optimization of matrix variables, avoiding dealing with the difficult rank constraint at the price of working with a highly non-convex objective function in U. Indeed, while the resulting optimization (4) has no rank constraint, a smaller memory footprint, and is more amenable for applying simple iterative algorithms like gradient descent, the permutational invariance of $f(U\circ U\circ U)$ implies that saddle points abound the optimization landscape among the exponentially many equivalent global minimizers. Unlike the original convex objective $f(T)$ that has an algorithm-friendly landscape where all the stationary points correspond to the global minimizers, the landscape for the resulting nonconvex formulation $f(U \circ U \circ U)$ is not well-understood. On the other hand, simple local search algorithms applied to (4) has exhibited superb empirical performance. As a first step towards understanding of the power of using the factorization method to solve tensor inverse problems, this work will focus on characterizing the local convergence of applying vanilla gradient descent to the general problem (4), as well as the global convergence of a simple variant.

Related Work

Burer-Monteiro Parameterization Method The idea of transforming the rank-constrained problem into an unconstrained problem using explicit factorization like $T = U \circ U \circ U$ is pioneered by Burer and Monteiro (Burer and Monteiro 2003, 2005) in solving matrix optimization problems with a rank constraint

$$\begin{aligned}&\underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}}\ f(X)\\ &\text{subject to}\ \ \mathrm{rank}(X)\le r \text{ and } X\succeq 0\end{aligned} \tag{5}$$

To deal with the rank constraint as well as the positive semidefinite constraint, the authors there proposed to firstly factorize a low-rank matrix $X = UU^\top$ with $U \in \mathbb{R}^{n\times r}$ and r chosen according to the rank constraint. Consequently, instead of minimizing an objective function $f(X)$ over all symmetric, positive semidefinite matrices of rank at most r, one can focus on an unconstrained nonconvex optimization:

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\, f(UU^\top).$$

Inspired by (Burer and Monteiro 2003, 2005), an intensive research effort has been devoted to investigating the theoretical properties of this factorization/parametrization method (Ge, Lee, and Ma 2016; Ge, Jin, and Zheng 2017; Park et al. 2017; Chi, Lu, and Chen 2019; Li, Zhu, and Tang 2018; Zhu et al. 2018, 2021; Li, Zhu, and Tang 2017; Zhu et al. 2019; Li et al. 2020). In particular, by analyzing the landscape of the resulting optimization, many authors have found that various low-rank matrix recovery problems in factored form–despite nonconvexity–enjoy a favorable landscape where all second-order stationary points are global minima.

Tensor Decomposition and Completion Another line of related work is nonconvex tensor factorization/completion. When the convex objective function $f(T)$ in (1) is the squared Euclidean distance between the tensor variable T and the ground-truth tensor $T^\star$, i.e., $f(T) = \|T - T^\star\|_F^2$, the resulting factorized problem (4) reduces to a (symmetric) tensor decomposition problem:

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\ f(U\circ U\circ U) = \|U\circ U\circ U - T^\star\|_F^2. \tag{6}$$

Tensor decomposition aims to identify the unknown rank-one factors from available tensor data. This problem is the backbone of several tensor-based machine learning methods, such as independent component analysis (Cardoso 1989) and collaborative filtering (Hou and Qian 2017). Unlike the similarly defined matrix decomposition, which has a closed-form solution given by the singular value decomposition, the tensor decomposition solution generally has no analytic expressions and is NP-hard to compute in the worst case (Hillar and Lim 2013). When the true tensor $T^\star$ is a fourth-order symmetric orthogonal tensor, i.e., there is an orthogonal matrix $U^\star$ such that $T^\star = U^\star \circ U^\star \circ U^\star \circ U^\star$, Ge et al. (Ge et al. 2015) designed a new objective function

$$\widetilde{f}(U) \doteq \sum_{i\neq j} \langle T^\star, \boldsymbol{u}_i \otimes \boldsymbol{u}_i \otimes \boldsymbol{u}_j \otimes \boldsymbol{u}_j\rangle$$

and showed that, despite its non-convexity, the objective function $\widetilde{f}(U)$ has a benign landscape on the sphere where all the local minima are global minima and all the saddle points have a Hessian with at least one negative eigenvalue. Later, (Qu et al. 2019) relax the orthogonal condition to near-orthogonal condition, resulting to landscape analysis to fourth-order overcomplete tensor decomposition. The work (Ge et al. 2015) has spurred many followups that dedicate on the analysis of the nonconvex optimization landscape of many other problems (Ge, Lee, and Ma 2016; Ge, Jin, and Zheng 2017; Bhojanapalli, Neyshabur, and Srebro 2016; Park et al. 2017; Chi, Lu, and Chen 2019). The techniques developed in (Ge et al. 2015), however, are not directly applicable to solve the original rank-constrained tensor optimization problem (6). In addition, (Ge et al. 2015) mainly considered fourth-order tensor decomposition, which cannot be trivially extended to analyze other odd-order tensor decompositions. More recently, Ge and Ma (Ge and Ma 2017) studied the problem of maximizing

$$\widehat{f}(\boldsymbol{u}) = \langle T, \boldsymbol{u}\otimes\boldsymbol{u}\otimes\boldsymbol{u}\otimes\boldsymbol{u}\rangle$$

on the unit sphere and presented a local convergence of applying vanilla gradient descent to the problem. Although this formulation together with iterative rank-1 updates lead to

algorithms with convergence guarantees for tensor decomposition, it is not flexible enough to deal with general rank-constrained problem (1). Similar rank-1 updating methods for tensor decomposition have also been investigated in (Anandkumar, Ge, and Janzamin 2017, 2015, 2014; Anandkumar et al. 2014).

More recently, (Chi, Lu, and Chen 2019; Cai et al. 2021) apply the factorization formulation to the tensor completion problem and focuses on solving

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}} \ \|P_\Omega\left(U\circ U\circ U-T^\star\right)\|_F^2, \tag{7}$$

where P_Ω is the the orthogonal projection of any tensor T onto the subspace indexed by the observation set Ω. (Chi, Lu, and Chen 2019; Cai et al. 2021) proposed a vanilla gradient descent following a rough initialization and proved the vanilla gradient descent could faithfully complete the tensor and retrieve all individual tensor factors within nearly linear time when the rank r does not exceed $O(n^{1/6})$. Compared with these prior state of the arts, our convergence analysis improves the order of rank r and extends the focus to general cost functions.

Main Contributions and Organization

To solve the rank-constrained tensor optimization problem (1), we directly work with the Burer-Monteiro factorized formulation (4) with a general convex function $f(\cdot)$ and focus on solving (4) using (vanilla) gradient descent

$$U^+ = U - \eta\nabla_U f(U\circ U\circ U), \tag{8}$$

where U^+ is the updated version of the current variable U, η is the stepsize that will be carefully tuned to prevent gradient descent from diverging, and $\nabla_U f$ is the gradient of $f(U\circ U\circ U)$ with respect to U.

In this work, we show that the factorized tensor minimization problem (4) satisfies the local regularity condition under certain mild assumptions. With this local regularity condition, we further prove a linear convergence of the gradient descent algorithm in a neighborhood of true tensor factors. In particular, we have shown that solving the factored tensor minimization problem (4) with gradient descent (8) is guaranteed to identify the target tensor $T^\star$ with high probability if $r = O\big(n^{1.25}\big)$ and n is sufficiently large. This implies that we can even deal with the scenario where the rank of the target tensor $T^\star$ is larger than the individual tensor dimensions, the so called overcomplete regime that are considered challenging to tackle in practice.

Finally, as a complement to the local analysis, we study the global landscape of best rank-1 approximation of a third-order orthogonal tensor and we show that this problem has no spurious local minima and all saddle points are strict saddle points except for the one at zero, which is a third-order saddle point.

Organization The remainder of this work is organized as follows. In Section 2, we first briefly introduce some basic definitions and concepts used in tensor analysis and then present the local convergence of applying vanilla gradient descent to the tensor minimization problem (4) and provide a linear convergence analysis for the gradient descent algorithm (8). In Section 3, we switch to analyze the global landscape of orthogonal tensor decomposition. Numerical simulations are conducted in Section 4 to further support our theory. Finally, we conclude our work in Section 5.

2 Local Convergence

In this section, we first briefly review some fundamental concepts and definitions in tensor analysis. A tensor with order higher than 3 can be viewed as a high-dimensional extension of vectors and matrices. In this work, we mainly focus on the third-order symmetric tensors. Any such tensor admits symmetric rank-one decompositions of the following form:

$$T = \sum_{p=1}^{r} c_p \boldsymbol{u}_p\otimes\boldsymbol{u}_p\otimes\boldsymbol{u}_p \in \mathbb{R}^{n\times n\times n}$$

with $\|\boldsymbol{u}_p\|_2 = 1$ and $c_p > 0$, $1\le p\le r$. The above decomposition is also called the *Canonical Polyadic* (CP) decomposition of the tensor T (Hong, Kolda, and Duersch 2020). The minimal number of factors r is defined as the *(symmetric) rank* of the tensor T. Denote $T(i_1,i_2,i_3)$ as the (i_1,i_2,i_3)-th entry of a tensor T. We define the *inner product* of any two tensors $X, Y\in\mathbb{R}^{n\times n\times n}$ as $\langle X,Y\rangle \doteq \sum_{i_1,i_2,i_3=1}^{n} X(i_1,i_2,i_3)Y(i_1,i_2,i_3)$. The induced *Frobenius norm* of a tensor T is then defined as $\|T\|_F \doteq \sqrt{\langle T,T\rangle}$. For a tensor $T\in\mathbb{R}^{n\times n\times n}$, we denote its unfolding/matricization along the first dimension as $T_{(1)} = [T(:,1,1)\ T(:,2,1)\ \cdots\ T(:,n,n)]\in\mathbb{R}^{n\times n^2}$.

We proceed to present the local convergence of applying vanilla gradient descent to the factored tensor minimization problem (4). Before that, we introduce several definitions used throughout the work.

Definition 1. *A function $f:\mathbb{R}^{n\times n\times n}\to\mathbb{R}$ is (r,m,M)-restricted strongly convex and smooth if*

$$m\|Y-X\|_F \le \|\nabla f(Y)-\nabla f(X)\|_F \le M\|Y-X\|_F$$

holds for any symmetric tensors $X, Y\in\mathbb{R}^{n\times n\times n}$ of rank at most r with some positive constants m and M.

For example, $f(T) = \frac{1}{2}\|T-T^\star\|_F^2$ is such a (r,m,M)-restricted strongly convex and smooth function for arbitrary $r\in\mathbb{N}$ with $M = m = 1$, and its global minimizer is $T = T^\star$.

Definition 2. *The distance between two factored matrices U_1 and U_2 is defined as*

$$\operatorname{dist}(U_1,U_2) = \min_{\text{Permutation}P}\|U_1-U_2P\|_F.$$

Denote

$$P_{U_1} = \arg\min_{\text{Permutation}P}\|U_1-U_2P\|_F. \tag{9}$$

Then, we can rewrite the distance between U_1 and U_2 as

$$\operatorname{dist}(U_1,U_2) = \|U_1-U_2P_{U_1}\|_F. \tag{10}$$

Define $\gamma\doteq\operatorname{polylog}(n)$ that may vary from place to place and $\widehat{U}\doteq\big[\widehat{\boldsymbol{u}}_1\ \widehat{\boldsymbol{u}}_2\ \cdots\ \widehat{\boldsymbol{u}}_r\big]$. Denote $\underline{c}\doteq\min_{p\in[r]}c_p^{\star 1/3}$, $\bar{c}\doteq\max_{p\in[r]}c_p^{\star 1/3}$, and $\omega = \bar{c}/\underline{c}$. We are ready to introduce the assumptions needed to prove our main theorem as follows.

Assumption 1. *(Incoherence condition). The vector factors $\widehat{\boldsymbol{u}}$ in the target tensor $T^\star$ satisfy*

$$\max_{i\neq j} |\langle \widehat{\boldsymbol{u}}_i, \widehat{\boldsymbol{u}}_j\rangle| \leq \frac{\gamma}{\sqrt{n}}.$$

Assumption 2. *(Bounded spectrum). The spectral norm of $\widehat{U}$ is bounded above as*

$$\|\widehat{U}\| \leq 1 + c_1\sqrt{\frac{r}{n}}.$$

Assumption 3. *(Isometry of Gram-matrix). The Gram matrix satisfies the following isometry property*

$$\|(\widehat{U}^\top\widehat{U}) \odot (\widehat{U}^\top\widehat{U}) - \mathbf{I}_r\| \leq \frac{\gamma\sqrt{r}}{n}.$$

where $\odot$ is the Hadamard product.

Assumption 4. *(Good current). The distance between the current variable U and the matrix factor $U^\star$ is bounded with*

$$\mathrm{dist}(U, U^\star) \leq 0.07\frac{m}{M}\frac{\underline{c}}{\omega^3}.$$

We remark that Assumptions 1-3 hold with high probability if the factors $\{\widehat{\boldsymbol{u}}_p\}_{p=1}^r$ are generated independently according to the uniform distribution on the unit sphere (Anandkumar, Ge, and Janzamin 2015, Lemmas 25, 31).

Main Results

We now present our main theorem in the following:

Theorem 1. *Suppose that a (r, m, M)-restricted strongly convex and smooth function $f : \mathbb{R}^{n\times n\times n} \to \mathbb{R}$ has a unique global minimizer at $T^\star$, which admits a CP decomposition $T^\star = U^\star \circ U^\star \circ U^\star \in \mathbb{R}^{n\times n\times n}$ as given in* (3)*. Then, under Assumptions 1-4 and in addition assuming $r = O\big(n^{1.25}\big)$, the following local regularity condition holds for sufficiently large n:*

$$\langle \nabla_U f(U\circ U\circ U), U - U^\star P_U\rangle \geq \frac{1}{2}\eta\|\nabla_U f(U\circ U\circ U)\|_F^2 + 0.13m\underline{c}^4\mathrm{dist}(U, U^\star)^2, \tag{11}$$

as long as

$$\eta \leq \frac{1}{18\|[\nabla f(T)]_{(1)}\|\cdot\|U\| + 9M\|U\|^4}. \tag{12}$$

Here $T = U\circ U\circ U$ and $[\nabla f(T)]_{(1)}$ denotes the matricization of $\nabla f(T)$ along the first dimension.

The local regularity condition further implies linear convergence of the gradient descent algorithm (8) in a neighborhood of the true tensor factors $U^\star$ with proper choice of the stepsize, as summarized in the following two corollaries.

Corollary 1 (Linear Convergence with adaptive stepsize)**.** *Under the same assumptions as in Theorem 1, we have the following (adaptive) linear convergence*

$$\begin{aligned}\mathrm{dist}(U^+, U^\star)^2 &\leq (1 - 0.26\eta m\underline{c}^4)\mathrm{dist}(U, U^\star)^2\\ &\doteq \alpha(\eta)\cdot\mathrm{dist}(U, U^\star)^2\end{aligned} \tag{13}$$

when we run the gradient descent algorithm (8) *with the stepsize η satisfying* (12)*.*

Corollary 2 (Linear Convergence with constant stepsize)**.** *Under the same assumptions as in Theorem 1, except that Assumption 4 is replaced by a good initial condition:*

$$\mathrm{dist}(U^0, U^\star) \leq 0.07\frac{m}{M}\frac{\underline{c}}{\omega^3}, \tag{14}$$

and the stepsize selection requirement (12) *is replaced with the constant stepsize satisfying $\eta_0 = \frac{1}{21.6M\|U^0\|^4}$, the sequence $\{U^t : t = 0, 1, 2, \cdots\}$ generated by*

$$U^{t+1} = U^t - \eta_0\nabla f(U^t\circ U^t\circ U^t),\ \ t = 0, 1, 2, \cdots$$

satisfies

$$\mathrm{dist}(U^+, U^\star)^2 \leq \alpha(\eta_0)\cdot\mathrm{dist}(U, U^\star)^2 \tag{15}$$

with $\alpha(\eta_0) \doteq 1 - 0.26\eta_0 m\underline{c}^4$.

As a consequence, we conclude that solving the factored problem (4) using the gradient descent algorithm (8) with a good initialization is guaranteed to recover the tensor factor matrix $U^\star$ with high probability if $r = O\big(n^{1.25}\big)$. The proof of the above theorem and corollaries can be found in supplementary material.

3 Global Convergence

The local convergence analysis of applying vanilla gradient descent to tensor optimization, though developed for a class of sufficiently general problems, is not completely satisfactory as a good initialization might be difficult to find. Therefore, we are also interested in characterizing the global optimization landscape for these problems. Considering the difficulty of this task, we focus on a special case where the ground-truth third-order tensor admits an orthogonal decomposition and we are interested in finding its best rank-one approximation. We aim to characterize all its critical points and classify them into local minima, strict saddle points, and degenerate saddle points if there is any. We also want to exploit the properties of critical points to design a provable and efficient tensor decomposition algorithm.

Main Results

Consider the best rank-one approximation problem of an orthogonally decomposable tensor:

$$g(\boldsymbol{u}) = \|\boldsymbol{u}\otimes\boldsymbol{u}\otimes\boldsymbol{u} - T^\star\|_F^2, \tag{16}$$

where $T^\star = \sum_{i=1}^r \boldsymbol{u}_i^\star\otimes\boldsymbol{u}_i^\star\otimes\boldsymbol{u}_i^\star$ and these true tensor factors $\{\boldsymbol{u}_i^\star\}$ are orthogonal to each other. This is a special case of (4) (and (6)). We characterize all possible critical points and their geometric properties in the following theorem:

Theorem 2. *Assume $T^\star = \sum_{i=1}^r \boldsymbol{u}_i^\star\otimes\boldsymbol{u}_i^\star\otimes\boldsymbol{u}_i^\star$, where $\{\boldsymbol{u}_i^\star\}$ are orthogonal to each other. Then any critical point $\widehat{\boldsymbol{u}}$ of $g(\boldsymbol{u})$ in* (16) *takes the form $\widehat{\boldsymbol{u}} = \sum_{i=1}^r \lambda_i\boldsymbol{u}_i^\star$ for $\boldsymbol{\lambda} \doteq [\lambda_1\ \cdots\ \lambda_r] \in \mathbb{R}^r$ and*

1. *when $\|\boldsymbol{\lambda}\|_0 = 0$, $\widehat{\boldsymbol{u}} = \mathbf{0}$ is a third-order saddle point, i.e., $\nabla^2 g(\widehat{\boldsymbol{u}}) = \mathbf{0}$ and $\nabla^3 g(\widehat{\boldsymbol{u}}) \neq \mathbf{0}$;*
2. *when $\|\boldsymbol{\lambda}\|_0 = 1$, $\widehat{\boldsymbol{u}} = \boldsymbol{u}_i^\star$ with $i \in \{1, 2, \ldots, r\}$ is a strict local minimum;*
3. *when $\|\boldsymbol{\lambda}\|_0 \geq 2$, $\widehat{\boldsymbol{u}}$ is a strict saddle point, i.e., $\nabla^2 g(\boldsymbol{u})$ has a negative eigenvalue.*

Algorithm 1: Iterative Gradient Descent for Tensor Decomposition

Input: $T^\star$
Initialization: $T = T^\star$, $\widehat{\boldsymbol{u}} = \mathbf{0}$
Output: Estimated factors $\{\boldsymbol{u}_i^\star\}$

Let $i = 0$.
while $T \neq \mathbf{0}$ **do**
  if $\widehat{\boldsymbol{u}} \neq \mathbf{0}$ **then**
    $i = i + 1$.
    $\boldsymbol{u}_i^\star = \widehat{\boldsymbol{u}}$
    $T \leftarrow T - \frac{\langle T, \widehat{\boldsymbol{u}}\otimes\widehat{\boldsymbol{u}}\otimes\widehat{\boldsymbol{u}}\rangle \widehat{\boldsymbol{u}}\otimes\widehat{\boldsymbol{u}}\otimes\widehat{\boldsymbol{u}}}{\|\widehat{\boldsymbol{u}}\|_2^3}$
  end if
  Find second-order stationary point $\widehat{\boldsymbol{u}}$ of $\|\boldsymbol{u}\otimes\boldsymbol{u}\otimes\boldsymbol{u} - T\|_F^2$.
end while
return solution

Here the ℓ_0 "norm" $\|\cdot\|_0$ counts the number of non-zero entries in a vector. Analytic expression for $\boldsymbol{\lambda}$ is given in the proof.

Theorem 2 implies that all second-order critical points are the true tensor factors except for zero. Based on this, we develop a provable conceptual tensor decomposition algorithm as follows:

Corollary 3. *Assume $T^\star$ is a third-order orthogonal tensor with the tensor factors $\{\boldsymbol{u}_i^\star\}$. Then with the input $T^\star$, Algorithm 1 almost surely recovers all the tensor factors $\{\boldsymbol{u}_i^\star\}$.*

Proof of Corollary 3. It mainly follows from the many iterative algorithms can find a second-order stationary point (Lee et al. 2016; Li, Zhu, and Tang 2019; Li et al. 2019; Nesterov and Polyak 2006; Jin et al. 2017). Then by Theorem 2, applying these iterative algorithms to $g(\boldsymbol{u})$, it converges to either a true tensor factor $\boldsymbol{u}_i^\star$ for $i \in [r]$ or the zero point (as a third-order saddle point is essentially a second-order stationary point). If it converges to a nonzero point, it must be a true tensor factor and we record it. Then we can remove this component by projecting the target tensor T into the orthogonal complement of $\boldsymbol{u}_i^\star$. We repeat this process to the new deflated tensor until we get a zero deflated tensor. That means, we have found all the true factors $\{\boldsymbol{u}_i^\star\}$. □

Proof of Theorem 2

Recall that

$$T^\star = \sum_{i=1}^r \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \doteq \sum_{i=1}^r \lambda_i \widehat{\boldsymbol{u}}_i \otimes \widehat{\boldsymbol{u}}_i \otimes \widehat{\boldsymbol{u}}_i.$$

Without loss of generality, we can extend the orthonormal set $\{\widehat{\boldsymbol{u}}_i\}_{i=1}^r$ to $\{\widehat{\boldsymbol{u}}_i\}_{i=1}^n$ as a full orthonormal basis of $\mathbb{R}^n$ and define

$$\lambda_i \doteq 0,\ i \in [r]^c \doteq \{r+1, \ldots, n\}.$$

Then, we have $T^\star = \sum_{i=1}^n \lambda_i \widehat{\boldsymbol{u}}_i \otimes \widehat{\boldsymbol{u}}_i \otimes \widehat{\boldsymbol{u}}_i$. Since $\{\widehat{\boldsymbol{u}}_i\}_{i=1}^n$ is a full orthonormal basis of $\mathbb{R}^n$, $\widehat{U} \doteq [\widehat{\boldsymbol{u}}_1 \cdots \widehat{\boldsymbol{u}}_n]$ is an orthonormal matrix, i.e., $\widehat{U}\widehat{U}^\top = \mathbf{I}$. Then the best rank-1 tensor approximation problem is equivalent to

$$\begin{aligned} g(\boldsymbol{u}) &= \|\boldsymbol{u}\otimes\boldsymbol{u}\otimes\boldsymbol{u} - T^\star\|_F^2 \\ &= \Big\|(\widehat{U}\widehat{U}^\top\boldsymbol{u})\otimes(\widehat{U}\widehat{U}^\top\boldsymbol{u})\otimes(\widehat{U}\widehat{U}^\top\boldsymbol{u}) \\ &\qquad - \sum_{i=1}^n \lambda_i \widehat{\boldsymbol{u}}_i\otimes\widehat{\boldsymbol{u}}_i\otimes\widehat{\boldsymbol{u}}_i\Big\|_F^2. \end{aligned} \tag{17}$$

Expanding the squared norm and using the fact that $\widehat{U}$ is orthonormal, we get

$$\begin{aligned} g(\boldsymbol{u}) &= \|(\widehat{U}^\top\boldsymbol{u})\otimes(\widehat{U}^\top\boldsymbol{u})\otimes(\widehat{U}^\top\boldsymbol{u}) - \mathrm{diag}_3(\boldsymbol{\lambda})\|_F^2 \\ &\doteq \widehat{g}(\widehat{U}^\top\boldsymbol{u}) \end{aligned} \tag{18}$$

where we denote

$$\begin{aligned} \widehat{g}(\boldsymbol{u}) &\doteq \|\boldsymbol{u}\otimes\boldsymbol{u}\otimes\boldsymbol{u} - \mathrm{diag}_3(\boldsymbol{\lambda})\|_F^2, \\ \mathrm{diag}_3(\boldsymbol{\lambda}) &\doteq \sum_{i=1}^n \lambda_i \boldsymbol{e}_i\otimes\boldsymbol{e}_i\otimes\boldsymbol{e}_i. \end{aligned}$$

Lemma 1. *The landscape of $g(\boldsymbol{u})$ and $\widehat{g}(\boldsymbol{u})$ are rotationally equivalent: $\boldsymbol{u}$ is a first/second-order stationary point of g if and only if $\widehat{U}^\top\boldsymbol{u}$ is a first/second-order stationary point of $\widehat{g}$.*

Proof of Lemma 1. Since $g(\boldsymbol{u}) = \widehat{g}(\widehat{U}^\top\boldsymbol{u})$, by chain rule,

$$\begin{aligned} \nabla g(\boldsymbol{u}) &= \widehat{U}\nabla\widehat{g}(\widehat{U}^\top\boldsymbol{u}), \\ \nabla^2 g(\boldsymbol{u}) &= \widehat{U}\nabla^2\widehat{g}(\widehat{U}^\top\boldsymbol{u})\widehat{U}^\top. \end{aligned} \tag{19}$$

Then it directly follows from the definitions of first/second stationary points. □

Therefore by Lemma 1, to understand the landscape of $g(\boldsymbol{u})$, it suffices to study that of

$$\widehat{g}(\boldsymbol{u}) = \|\boldsymbol{u}\otimes\boldsymbol{u}\otimes\boldsymbol{u} - \mathrm{diag}_3(\boldsymbol{\lambda})\|_F^2.$$

We compute its derivatives up to third-order:

$$\begin{aligned} \nabla\widehat{g}(\boldsymbol{u}) &= 6\|\boldsymbol{u}\|_2^4\boldsymbol{u} - 6\boldsymbol{\lambda}\odot\boldsymbol{u}\odot\boldsymbol{u}, \\ \nabla^2\widehat{g}(\boldsymbol{u}) &= 6\|\boldsymbol{u}\|_2^4\mathbf{I} + 24\|\boldsymbol{u}\|_2^2\boldsymbol{u}\boldsymbol{u}^\top - 12\mathrm{diag}_3(\boldsymbol{\lambda}\odot\boldsymbol{u}), \\ \nabla^3\widehat{g}(\boldsymbol{u}) &= 24\|\boldsymbol{u}\|_2^2\mathrm{Sym}(\mathbf{I}\otimes\boldsymbol{u}) + 48\boldsymbol{u}\otimes\boldsymbol{u}\otimes\boldsymbol{u} - 12\mathrm{diag}_3(\boldsymbol{\lambda}), \end{aligned}$$

where $\odot$ is the Hadamard product and $\mathrm{Sym}(T)$ is the sum of all the three permutations of T.

Now define J as the index set of any critical point $\boldsymbol{u}$ such that $u_i \neq 0$ for $i \in J$, i.e.,

$$\boldsymbol{u}_J \neq \mathbf{0},\ \boldsymbol{u}_{J^c} = \mathbf{0},\ \|\boldsymbol{u}\|_0 = |J|.$$

By the critical point equation

$$\widehat{\boldsymbol{u}}\|_2^4\widehat{\boldsymbol{u}} - \boldsymbol{\lambda}\odot\widehat{\boldsymbol{u}}\odot\widehat{\boldsymbol{u}} = \mathbf{0} \tag{20}$$

and $\lambda_i = 0$ for $i \in [r]^c$, we conclude that $J \subset [r]$. In the following, we divide the problem into three cases: $|J| = 0, |J| = 1,$ and $|J| \geq 2$.

- Case I: $|J| = 0$. That is $\widehat{\boldsymbol{u}} = \mathbf{0}$. Then, we have

$$\nabla \widehat{g}(\widehat{\boldsymbol{u}}) = \mathbf{0}, \text{ and } \nabla^2 \widehat{g}(\widehat{\boldsymbol{u}}) = \mathbf{0},$$

but

$$\nabla^3 \widehat{g}(\widehat{\boldsymbol{u}}) = -12\mathrm{diag}_3(\boldsymbol{\lambda}).$$

This implies $\widehat{\boldsymbol{u}} = \mathbf{0}$ is a third-order saddle point of $\widehat{g}$.

- Case II: $|J| = 1$. Since $J \subset [r]$, let $J = \{k\}$ for some $k \in [r]$. Then,

$$\widehat{\boldsymbol{u}} = \widehat{u}_k \boldsymbol{e}_k \text{ and } \nabla \widehat{g}(\widehat{\boldsymbol{u}}) = 6\widehat{u}_k^5 \boldsymbol{e}_k - 6\lambda_k \widehat{u}_k^2 \boldsymbol{e}_k = \mathbf{0},$$

which implies that $\widehat{u}_k = \sqrt[3]{\lambda_k}$. We also have

$$\begin{aligned}\nabla^2 \widehat{g}(\widehat{\boldsymbol{u}}) &= 6\lambda_k^{4/3}\mathbf{I} + 24\lambda_k^{4/3}\boldsymbol{e}_k\boldsymbol{e}_k^\top - 12\lambda_k^{4/3}\boldsymbol{e}_k\boldsymbol{e}_k^\top \\ &= 6\lambda_k^{4/3}\mathbf{I} + 12\lambda_k^{4/3}\boldsymbol{e}_k\boldsymbol{e}_k^\top \succ 0\end{aligned}$$

Therefore, any critical point $\widehat{\boldsymbol{u}}$ with $\|\widehat{\boldsymbol{u}}\|_0 = 1$ has the form $\widehat{\boldsymbol{u}} = \sqrt[3]{\lambda_k}\boldsymbol{e}_k$ for $k \in [r]$, and is a strict local minimum of $\widehat{g}$.

- Case III: $|J| \geq 2$. Also, we know that $J \subset [r]$. With the critical point equation, we get

$$\boldsymbol{\lambda}_J \odot \widehat{\boldsymbol{u}}_J = \|\widehat{\boldsymbol{u}}\|_2^4 \mathbf{1}_{|J|}.$$

Further notice that

$$\mathrm{diag}_3(\boldsymbol{\lambda}_J \odot \widehat{\boldsymbol{u}}_J) = \|\widehat{\boldsymbol{u}}\|_2^4 \mathbf{I}_{|J|}.$$

Plugging this to the sub-Hessian

$$[\nabla^2 \widehat{g}(\widehat{\boldsymbol{u}})]_{J,J} = 24\|\widehat{\boldsymbol{u}}\|_2^2 \widehat{\boldsymbol{u}}_J \widehat{\boldsymbol{u}}_J^\top - 6\|\widehat{\boldsymbol{u}}\|_2^4 \mathbf{I}_{|J|}.$$

Now for any $\boldsymbol{d} \in \mathbb{R}^n$ with $\boldsymbol{d}_J^\top \boldsymbol{u}_J = 0, \boldsymbol{d}_{J^c} = \mathbf{0}$, we have

$$(\boldsymbol{d}, \boldsymbol{d}) = -6\|\widehat{\boldsymbol{u}}\|_2^4 \|\boldsymbol{d}_J\|_2^2 = -6\|\widehat{\boldsymbol{u}}\|_2^4 \|\boldsymbol{d}\|_2^2,$$

implying that

$$\lambda_{\min}(\nabla^2 \widehat{g}(\widehat{\boldsymbol{u}})) \leq -6\|\widehat{\boldsymbol{u}}\|_2^4 < 0.$$

Therefore, any critical point $\widehat{\boldsymbol{u}}$ with $\|\widehat{\boldsymbol{u}}\|_0 \geq 2$ has the form $\widehat{\boldsymbol{u}}_J = \frac{\|\widehat{\boldsymbol{u}}_J\|_2^4}{\boldsymbol{\lambda}_J}$ (pointwise), and is a strict saddle point of $\widehat{g}$.

Together with Lemma 1, we complete the proof of Theorem 2.

4 Numerical Experiments

Computing Infrastructure All the numerical experiments are performed on a 2018 MacBook Pro with operating system of macOS version 10.15.7, processor of 2.6 GHz 6-Core Intel Core i7, memory of 32 GB, and MATLAB version of R2020a.

In the first experiment, we illustrate the linear convergence of the gradient descent algorithm within the contraction region $\mathrm{dist}(U^0, U^\star) \leq 0.07\frac{m}{M}\frac{\underline{c}}{\omega^3}$ in solving the tensor decomposition problem (6), where $M = m = 1$ in this case. We set $n = 64$ and vary r with three different values: $n/2,\ n,\ 3n/2$ to get an undercomplete, complete, and overcomplete target tensor $T^\star$, respectively. We generate the r columns of $U^\star$ independently according to the uniform distribution on the unit sphere and form $T^\star = U^\star \circ U^\star \circ U^\star$. According to (Anandkumar, Ge, and Janzamin 2015, Lemmas 25, 31) and Corollary 2, if $\mathrm{dist}(U^0, U^\star) \leq 0.07\frac{m}{M}\frac{\underline{c}}{\omega^3} = 0.07$ (because $\|\boldsymbol{u}_i^\star\|_2 = 1$ implies $\bar{c} = \underline{c} = \omega = 1$), the gradient descent with a sufficiently small constant stepsize would converge linearly to the true factor $U^\star$. To illustrate this, we initialize the starting point as $U^\star + \alpha D$ with $\alpha = 0.07$ and set D as a normalized Gaussian matrix with $\|D\|_F = 1$. We record the three metrics $\|\nabla f(U)\|_F$, $\|U \circ U \circ U - T^\star\|_F$, and $\mathrm{dist}(U, U^\star)$ for total 10^3 iterations with different stepsizes η in Figure 1, which is consistent with the linear convergence analysis of gradient descent on general Burer-Monteiro tensor optimizations in Corollary 2.

In the second experiment, with the same settings as above except varying α, we record the success rate by running 100 trials for each fixed (r, α)-pair and declare one successful instance if the final iterate U satisfies $\mathrm{dist}(U, U^\star) \leq 10^{-3}$. We repeat these experiments for different $\alpha \in \{0.5, 1, 2, 4, 8, 16\}$. Table 1 shows that when α is small enough ($\alpha \leq 2$), the success rate is 100% for all the undercomplete ($r = n/2$), complete ($r = n$), and overcomplete ($r = 3n/2$) cases; and when α is comparatively large ($\alpha \in [4, 8]$), the success rate degrades dramatically when r increases. Finally, when α is larger than certain threshold, the success rate is 0%. This in consistence with Corollaries 1 and 2.

Table 1: Success ratio with $\eta = 0.02$ (*top*), $\eta = 0.04$ (*middle*), and $\eta = 0.06$ (*bottom*).

α	0.5	1	2	4	8	16
$r = n/2$	100%	100%	100%	100%	100%	0%
$r = n$	100%	100%	100%	100%	100%	0%
$r = 3n/2$	100%	100%	100%	100%	5%	0%

α	0.5	1	2	4	8	16
$r = n/2$	100%	100%	100%	100%	100%	0%
$r = n$	100%	100%	100%	100%	0%	0%
$r = 3n/2$	100%	100%	100%	100%	0%	0%

α	0.5	1	2	4	8	16
$r = n/2$	100%	100%	100%	100%	38%	0%
$r = n$	100%	100%	100%	100%	0%	0%
$r = 3n/2$	100%	100%	100%	83%	0%	0%

5 Conclusion

In this work, we investigated the local convergence of third-order tensor optimization with general convex and well-conditioned objective functions. Under certain incoherent conditions, we proved the local regularity condition for the nonconvex factored tensor optimization resulted from the Burer-Monteiro reparameterization. We highlighted that these assumptions are satisfied for randomly generated tensor factors. With this local regularity condition, we further provided a linear convergence analysis for the gradient descent algorithm started in a neighborhood of the true tensor factors. Complimentary to the local analysis, we also presented a complete characterization of the global optimization landscape of the best rank-one tensor approximation problem.

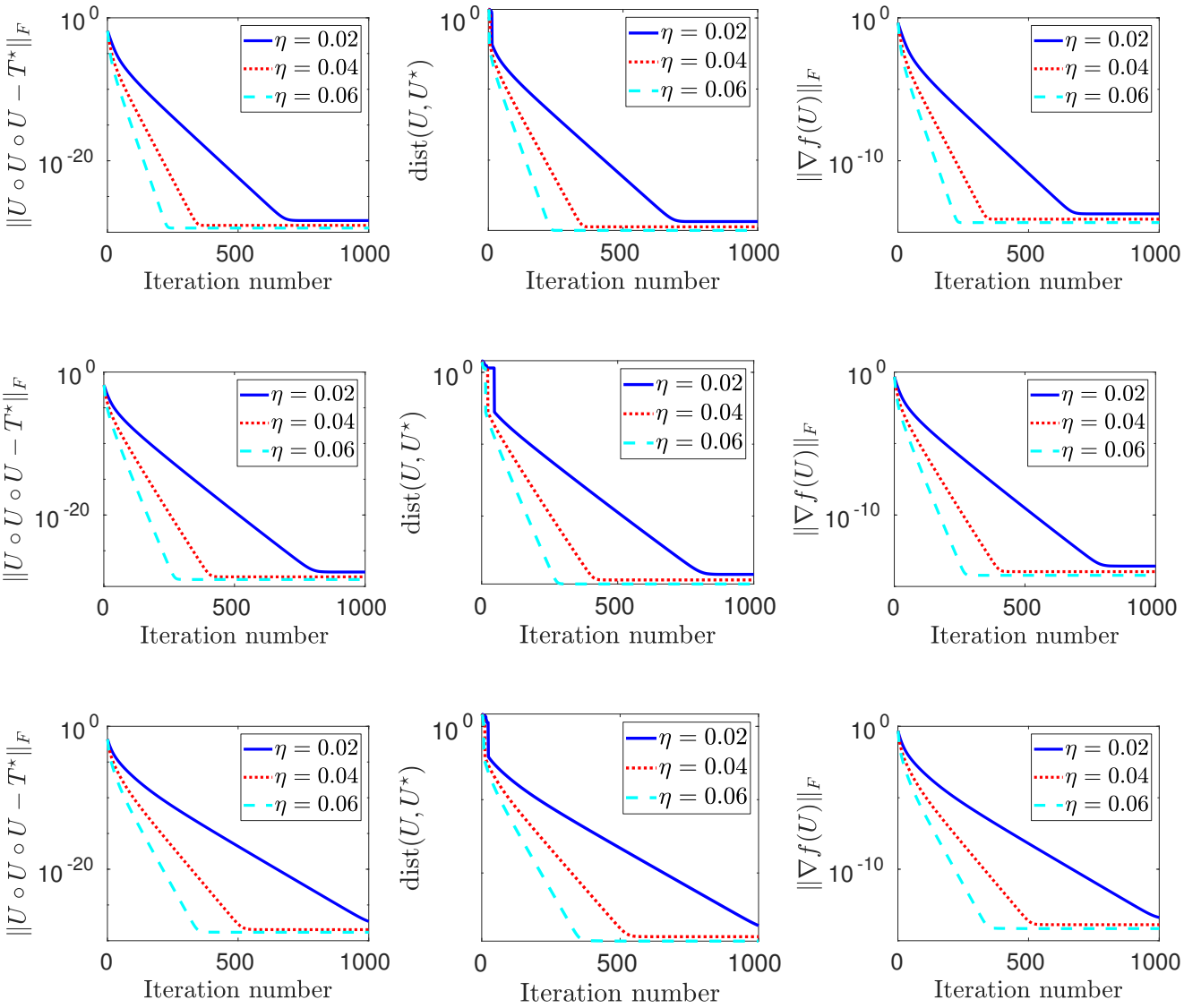

Figure 1: Linear convergence of gradient descent when applied to tensor factorization problem (6). Here, $r = n/2$ (*top row*), $r = n$ (*middle row*), and $r = 3n/2$ (*bottom row*) with $n = 64$. We initialize the starting point as $U^\star + \alpha D$ with $\alpha = 0.07$ and set D as a normalized Gaussian matrix with $\|D\|_F = 1$. We record the three metrics $\|\nabla f(U)\|_F$ (*left column*), $\|U \circ U \circ U - T^\star\|_F$ (*middle column*), and $\mathrm{dist}(U, U^\star)$ (*right column*) for total 10^3 iterations with different stepsize η, which is consistent with the linear convergence analysis of gradient descent on general Burer-Monteiro tensor optimizations.

Acknowledgments

S. Li and Q. Li would like to thank Prof. Gongguo Tang for many helpful discussions.

References

Anandkumar, A.; Ge, R.; Hsu, D.; Kakade, S. M.; and Telgarsky, M. 2014. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1): 2773–2832.

Anandkumar, A.; Ge, R.; and Janzamin, M. 2014. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. *arXiv preprint arXiv:1402.5180*.

Anandkumar, A.; Ge, R.; and Janzamin, M. 2015. Learning overcomplete latent variable models through tensor methods. In *Conference on Learning Theory*, 36–112.

Anandkumar, A.; Ge, R.; and Janzamin, M. 2017. Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research*, 18(22): 1–40.

Bai, Z.; Li, Y.; Woźniak, M.; Zhou, M.; and Li, D. 2021. Decomvqanet: Decomposing visual question answering deep network via tensor decomposition and regression. *Pattern Recognition*, 110: 107538.

Barak, B.; Kelner, J. A.; and Steurer, D. 2015. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, 143–151. ACM.

Barak, B.; and Moitra, A. 2016. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, 417–445. PMLR.

Bhojanapalli, S.; Neyshabur, B.; and Srebro, N. 2016. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, 3873–3881.

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

Burer, S.; and Monteiro, R. D. 2005. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3): 427–444.

Cai, C.; Li, G.; Poor, H. V.; and Chen, Y. 2021. Nonconvex low-rank tensor completion from noisy data. *Operations Research*.

Cardoso, J.-F. 1989. Source separation using higher order moments. *International Conference on Acoustics, Speech, and Signal Processing*, 2109–2112 vol.4.

Chi, Y.; Lu, Y. M.; and Chen, Y. 2019. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20): 5239–5269.

Ge, R.; Huang, F.; Jin, C.; and Yuan, Y. 2015. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, 797–842. PMLR.

Ge, R.; Jin, C.; and Zheng, Y. 2017. No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis. In *Proceedings of the 34th International Conference on Machine Learning*, 1233–1242. PMLR.

Ge, R.; Lee, J. D.; and Ma, T. 2016. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, 2973–2981.

Ge, R.; and Ma, T. 2017. On the optimization landscape of tensor decompositions. In *Advances in Neural Information Processing Systems*, 3653–3663.

Grasedyck, L.; Kressner, D.; and Tobler, C. 2013. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1): 53–78.

Hillar, C. J.; and Lim, L.-H. 2013. Most tensor problems are NP-Hard. *Journal of the ACM (JACM)*, 60(6): 45–39.

Hong, D.; Kolda, T. G.; and Duersch, J. A. 2020. Generalized canonical polyadic tensor decomposition. *SIAM Review*, 62(1): 133–163.

Hou, J.; and Qian, H. 2017. Collaboratively filtering malware infections: a tensor decomposition approach. In *Proceedings of the ACM Turing 50th Celebration Conference-China*, 28. ACM.

Jin, C.; Ge, R.; Netrapalli, P.; Kakade, S. M.; and Jordan, M. I. 2017. How to escape saddle points efficiently. In *International Conference on Machine Learning*, 1724–1732. PMLR.

Kolda, T. G.; and Bader, B. W. 2009. Tensor decompositions and applications. *SIAM Review*, 51(3): 455–500.

Lee, J. D.; Simchowitz, M.; Jordan, M. I.; and Recht, B. 2016. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, 1246–1257.

Li, Q.; Prater, A.; Shen, L.; and Tang, G. 2015a. Overcomplete tensor decomposition via convex optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 53–56. IEEE.

Li, Q.; Prater, A.; Shen, L.; and Tang, G. 2016. A super-resolution framework for tensor decomposition. *arXiv preprint arXiv:1602.08614*.

Li, Q.; and Tang, G. 2017. Convex and nonconvex geometries of symmetric tensor factorization. In *Asilomar Conference on Signals, Systems, and Computers*.

Li, Q.; Zhu, Z.; and Tang, G. 2017. Geometry of factored nuclear norm regularization. *arXiv preprint arXiv:1704.01265*.

Li, Q.; Zhu, Z.; and Tang, G. 2018. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1): 51–96.

Li, Q.; Zhu, Z.; and Tang, G. 2019. Alternating minimizations converge to second-order optimal solutions. In *International Conference on Machine Learning*, 3935–3943. PMLR.

Li, Q.; Zhu, Z.; Tang, G.; and Wakin, M. B. 2019. Provable bregman-divergence based methods for nonconvex and non-lipschitz problems. *arXiv preprint arXiv:1904.09712*.

Li, S.; Li, Q.; Zhu, Z.; Tang, G.; and Wakin, M. B. 2020. The global geometry of centralized and distributed low-rank matrix recovery without regularization. *IEEE Signal Processing Letters*, 27: 1400–1404.

Li, S.; Wang, W.; Qi, H.; Ayhan, B.; Kwan, C.; and Vance, S. 2015b. Low-rank tensor decomposition based anomaly detection for hyperspectral imagery. In *2015 IEEE International Conference on Image Processing (ICIP)*, 4525–4529. IEEE.

Liu, J.; Musialski, P.; Wonka, P.; and Ye, J. 2012. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 208–220.

Nesterov, Y.; and Polyak, B. T. 2006. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1): 177–205.

Park, D.; Kyrillidis, A.; Carmanis, C.; and Sanghavi, S. 2017. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, 65–74.

Phan, A.-H.; Sobolev, K.; Sozykin, K.; Ermilov, D.; Gusak, J.; Tichavskỳ, P.; Glukhov, V.; Oseledets, I.; and Cichocki, A. 2020. Stable low-rank tensor decomposition for compression of convolutional neural network. In *European Conference on Computer Vision*, 522–539. Springer.

Qu, Q.; Zhai, Y.; Li, X.; Zhang, Y.; and Zhu, Z. 2019. Geometric analysis of nonconvex optimization landscapes for overcomplete learning. In *International Conference on Learning Representations*.

Sedghi, H.; Janzamin, M.; and Anandkumar, A. 2016. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, 1223–1231.

Sidiropoulos, N. D.; De Lathauwer, L.; Fu, X.; Huang, K.; Papalexakis, E. E.; and Faloutsos, C. 2017. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13): 3551–3582.

Smilde, A.; Bro, R.; and Geladi, P. 2005. *Multi-Way Analysis: Applications in the Chemical Sciences*. John Wiley & Sons.

Sun, L.; Wu, F.; Zhan, T.; Liu, W.; Wang, J.; and Jeon, B. 2020. Weighted nonlocal low-rank tensor decomposition method for sparse unmixing of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13: 1174–1188.

Tang, G.; and Shah, P. 2015. Guaranteed tensor decomposition: A moment approach. In *International Conference on Machine Learning*, 1491–1500. PMLR.

Yuan, M.; and Zhang, C.-H. 2016. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4): 1031–1068.

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

Zhu, Z.; Li, Q.; Tang, G.; and Wakin, M. B. 2021. The global optimization geometry of low-rank matrix optimization. *IEEE Transactions on Information Theory*, 67(2): 1308–1331.

Zhu, Z.; Li, Q.; Yang, X.; Tang, G.; and Wakin, M. B. 2019. Distributed Low-rank Matrix Factorization With Exact Consensus. *Advances in Neural Information Processing Systems*, 32: 8422–8432.