Ann. Appl. Math.
doi: 10.4208/aam.OA-2021-0090

Vol. **38**, No. 4, pp. 385-413
November 2022

Gradient Descent for Symmetric Tensor Decomposition

Jian-Feng Cai[1], Haixia Liu[2,*] and Yang Wang[1]

[1] *Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China*
[2] *School of Mathematics and Statistics & Hubei Key Laboratory of Engineering Modeling and Scientific Computing, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*

Received 13 December 2021; Accepted (in revised version) 30 August 2022

Abstract. Symmetric tensor decomposition is of great importance in applications. Several studies have employed a greedy approach, where the main idea is to first find a best rank-one approximation of a given tensor, and then repeat the process to the residual tensor by subtracting the rank-one component. In this paper, we focus on finding a best rank-one approximation of a given orthogonally order-3 symmetric tensor. We give a geometric landscape analysis of a nonconvex optimization for the best rank-one approximation of orthogonally symmetric tensors. We show that any local minimizer must be a factor in this orthogonally symmetric tensor decomposition, and any other critical points are linear combinations of the factors. Then, we propose a gradient descent algorithm with a carefully designed initialization to solve this nonconvex optimization problem, and we prove that the algorithm converges to the global minimum with high probability for orthogonal decomposable tensors. This result, combined with the landscape analysis, reveals that the greedy algorithm will get the tensor CP low-rank decomposition. Numerical results are provided to verify our theoretical results.

AMS subject classifications: 65F15, 93B60

Key words: Gradient descent, random initialization, symmetric tensor decomposition, CP decomposition, linear convergence.

*Corresponding author.
Emails: jfcai@ust.hk (J. Cai), liuhaixia@hust.edu.cn (H. Liu), yangwang@ust.hk (Y. Wang)

1 Introduction

Tensor decomposition can be viewed as an extension of the singular value decomposition (SVD) for matrices, which is obviously one of the fundamental tools in numerous applications. Unlike for matrices, the term "decomposition" for tensors can carry very different meanings in different studies. In this paper we focus on one of the most commonly used notions of tensor decomposition: the *canonical polyadic decomposition (CP decomposition, or CPD).*

Before going further we first introduce some notations. Let $\mathcal{A}$ be a tensor, which is an element of $\bigotimes_{j=1}^{m}\mathbb{R}^{n_j}:=\mathbb{R}^{n_1\times n_2\times\cdots\times n_m}$. A *rank-one* tensor $\mathcal{A}$ in $\bigotimes_{j=1}^{m}\mathbb{R}^{n_j}$ has the form

$$\mathcal{A}=\prod_{k=1}^{m}\mathbf{v}_k:=\mathbf{v}_1\otimes\mathbf{v}_2\otimes\cdots\otimes\mathbf{v}_m,$$

namely $[\mathcal{A}]_{j_1j_2\cdots j_m}=v_{1j_1}v_{2j_2}\cdots v_{mj_m}$. For simplicity we use $\mathbf{j}$ to denote the multi-index $\mathbf{j}:=(j_1,j_2,\cdots,j_m)$, and $[\mathcal{A}]_{\mathbf{j}}$ to denote the $\mathbf{j}$-th entry of $\mathcal{A}$. Given a general tensor $\mathcal{A}\in\bigotimes_{j=1}^{m}\mathbb{R}^{n_j}$, a *CP decomposition (CPD)* of $\mathcal{A}$ is to decompose it into sum of rank-one tensors, $\mathcal{A}=\mathcal{A}_1+\mathcal{A}_2+\cdots+\mathcal{A}_r$, where each $\mathcal{A}_i$, $i=1,\cdots,r$ is a rank-one tensor. The minimal r is called the *CP rank* of $\mathcal{A}$. A major problem in tensor decomposition is to compute the CP rank and the CP decomposition of a tensor.

A particularly important class of tensors are the so-called super symmetric tensors, or simply *symmetric tensors*. A tensor $\mathcal{A}\in\bigotimes^{m}\mathbb{R}^{n}$ is called a symmetric tensor if for any multi-index $\mathbf{j}\in\{1,2,\cdots,n\}^m$ and any permutation $\mathbf{i}$ of $\mathbf{j}$ we have $[\mathcal{A}]_{\mathbf{i}}=[\mathcal{A}]_{\mathbf{j}}$. It is easy to see that a rank-one symmetric tensor $\mathcal{A}$ must have the form

$$\mathcal{A}=\lambda\mathbf{v}^{\otimes m}:=\lambda\underbrace{\mathbf{v}\otimes\cdots\otimes\mathbf{v}}_{m},$$

where $\mathbf{v}\in\mathbb{R}^n$ and $\lambda\in\mathbb{R}$ are both nonzero. For a general symmetric tensor $\mathcal{A}$, a *symmetric CP decomposition (symmetric CPD)* is

$$\mathcal{A}=\mathcal{A}_1+\mathcal{A}_2+\cdots+\mathcal{A}_r,$$

where each $\mathcal{A}_i$, $i=1,\cdots,r$ is a symmetric rank-one tensor. The minimal r is called the *symmetric CP rank* of $\mathcal{A}$. Like general tensor decomposition, symmetric CP decomposition of a symmetric tensor is a major (and challenging) problem in the study of tensors. Although finding the symmetric CP rank of a symmetric tensor and its symmetric CP decomposition are generally very challenging, they are very useful in applications. For example, symmetric tensors appear as higher order derivatives or moments and cumulants of random vectors, which are often used in source extraction, mobile communications, machine learning, factor analysis of m-way arrays, biomedical engineering, psychometrics, and chemometrics [4,6–9,16,25].

More recently, ℓ^p norm with $p>2$ maximization [24, 27] is introduced instead of ℓ^1 norm minimization to exploit sparsity and learn complete dictionary, which can be considered as a CP orthogonal decomposition problem.

A general formulation is to convert them into optimization problems. Let $\mathcal{A}$ be a symmetric tensor. Then for any $s>0$ the *best symmetric CP rank-s approximation* of $\mathcal{A}$ is to solve the minimization problem

$$\min_{\lambda_j,\|\mathbf{v}_j\|=1} \left\|\mathcal{A}-\sum_{j=1}^{s}\lambda_j\mathbf{v}_j^{\otimes m}\right\|_F^2, \tag{1.1}$$

where $\lambda_j\in\mathbb{R}$, $\mathbf{v}_j\in\mathbb{R}^n$ and the Frobenious norm of a tensor $\mathcal{A}$ is defined as

$$\|\mathcal{A}\|_F := \sqrt{\sum_{\mathbf{i}\in[n]^m}[\mathcal{A}]_{\mathbf{i}}^2} \quad \text{with} \quad [n]=\{1,2,\cdots,n\}$$

for simplicity. Unfortunately, the above optimization problem (1.1) is not convex. Note that in the case of real field with $m=2$, the problem (1.1) can be solved through the greedy algorithm in the sense that we first find the largest singular value and singular vector, subtract the corresponding component from $\mathcal{A}$, and repeat the process. This greedy algorithm guarantees that we actually get the optimal best rank r approximation of $\mathcal{A}$. Several studies have employed this approach for tensors with order $m>2$ [1, 2, 10–12, 14, 15, 18, 19, 21–23]. In this approach, one first obtain the best symmetric rank-1 decomposition by solving the optimization problem

$$\min_{\lambda,\|\mathbf{x}\|=1} \frac{1}{2}\left\|\mathcal{A}-\lambda\mathbf{x}^{\otimes m}\right\|_F^2, \tag{1.2}$$

which is equivalent to

$$\max |\langle\mathcal{A},\mathbf{x}^{\otimes m}\rangle| \quad \text{subject to} \quad \|\mathbf{x}\|=1, \tag{1.3}$$

where the inner product is

$$\langle\mathcal{A},\mathcal{B}\rangle = \sum_{\mathbf{i}\in[n]^m}[\mathcal{A}]_{\mathbf{i}}[\mathcal{B}]_{\mathbf{i}}.$$

Once the best symmetric rank-one approximation is obtained, we repeat the procedure on the residue tensor $\mathcal{A}-\lambda\mathbf{x}^{\otimes m}$. After r iterations we find a rank r approximation of the symmetric tensor $\mathcal{A}$.

Problem (1.3) is unfortunately an NP-hard problem when $m>2$ [11]. A variety of methods have been introduced to solve it, see e.g., [11, 12, 14, 15, 18, 21, 22]. Qi,

Wang and Wang [22] proposed some Z-eigenvalue methods for solving the symmetric CP decomposition problem. Kofidis and Regalia [12] consider the high-order power method (HOPM) and explain the condition under which the method is convergent for even-order symmetric tensors. A shifted symmetric higher-order power method (SS-HOPM) [15] is introduced by Kolda et al. for computing tensor eigenpairs and give the convergence guarantee to a tensor eigenpair. Anandkumar et al. [1] give a robust power method of tensor decompositions for learning latent variable models. They provide detailed perturbation analysis for a robust variant of tensor power method. Although the convergence rate for orthogonally decomposable tensors is quadratic, the optimizer relies on the choice of initialization. Kolda [13] points out there is a transformation from non-orthogonal tensor decomposition to orthogonal tensor decomposition when the matrix $V=[\mathbf{v}_1,\cdots,\mathbf{v}_r]$ is with full column rank. Jiang et al. [11] introduced tensor principal component analysis (tensor PCA) via convex optimization. They showed the general tensor PCA problem is reducible when the tensor is supersymmetric with an even order. They prove that if the tensor is rank-1, then the embedded matrix must be rank-1, and vice versa. To enforce a low-rank solution, they impose a nuclear norm penalty or relax the rank-1 constraint by semidefinite programming. To further cope with the size of the resulting convex optimization models, they solve by ADMM. Liu uses alternating minimization to find CP decomposition of symmetric tensors [19]. Pan et al. [20] apply the symmetric orthogonal approximation to symmetric tensor to image reconstructions, which is tensor power method plus orthogonal projection.

Despite numerous aforementioned efficient best symmetric rank-one approximation algorithms by solving (1.2) or its variants, no algorithm is theoretically guaranteed to find the maximal component in obtaining the best rank-1 decomposition. We aims at filling this gap in this paper. We mainly focus on rank-r orthogonally decomposible symmetric tensors of order-3. In this case, to find the best rank-one approximation of a symmetric tensor $\mathcal{A}\in\mathbb{R}^{\otimes 3}$, the nonconvex least squares model (1.2) is equivalent to the following

$$\min_{\mathbf{z}} f(\mathbf{z}), \quad \text{where} \quad f(\mathbf{z})=\frac{1}{6}\|\mathcal{A}-\mathbf{z}^{\otimes 3}\|_F^2. \tag{1.4}$$

The contribution of this paper is two-folded. Firstly, we give a geometric landscape analysis of the nonconvex function f in Section 2. In particular, we show that any local minimizer must be a factor in the CP low-rank decomposition of $\mathcal{A}$, and any other critical points are linear combinations of the factors. Then, we propose in Section 3 a gradient descent algorithm with a well-designed initialization to solve (1.4), and prove that the algorithm converges to the global minimizer (i.e., the best rank-one approximation) with high probability. This result, combined with the

landscape of f, reveals that the greedy algorithm with carefully initialized gradient descent get the CP low-rank decomposition of $\mathcal{A}$.

2 Landscape of nonconvex optimization model to solve tensor decomposition

In this section, we analyze the landscapes of the nonconvex function f in (1.4). That is, the locations and the characterization of local/global optimum and saddle points. We show that, for an orthogonally symmetric decomposable tensor $\mathcal{A}$, the factors in the CP low-rank decomposition of $\mathcal{A}$ are only local minimizers of (1.4), and all other non-zero critical points are linear combinations of the factors, which are strict saddle points. The result is presented in Theorem 2.1. A similar result was presented in [1].

Theorem 2.1 (Classification of Critical Points)**.** *Let*

$$\mathcal{A}=\sum_{i=1}^{r}\mathbf{x}_i\otimes\mathbf{x}_i\otimes\mathbf{x}_i,$$

where $\{\mathbf{x}_i\}_{i=1}^r$ *are nonzero orthogonal vectors and* $\|\mathbf{x}_1\|\geq\|\mathbf{x}_2\|\geq\cdots\geq\|\mathbf{x}_r\|>0$*. Then any critical point* $\mathbf{z}$ *of* f *defined in* (1.4) *is in the form of*

$$\mathbf{z}=\sum_{j\in\mathcal{I}}\frac{\|\mathbf{z}\|^4}{\|\mathbf{x}_j\|^4}\mathbf{x}_j,\qquad \mathcal{I}\subseteq[r].$$

Moreover, we have

1. *When* $|\mathcal{I}|=1$*, the critical points are* $\mathbf{x}_j$*,* $j=1,\cdots,r$*, which are all local minimizers.*
2. *When* $|\mathcal{I}|\geq 2$*, the critical points are strict saddle points.*
3. *When* $|\mathcal{I}|=0$*, the critical point is* $\mathbf{0}$*, which is not a local minimizer.*

Here $|\mathcal{I}|$ *stands for the cardinality of a set* $\mathcal{I}$*.*

Proof. We expand $\{\mathbf{x}_1,\cdots,\mathbf{x}_r\}$ to be an orthogonal (not necessarily normal) basis of $\mathbb{R}^n$, denoted by $\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$, where all vectors are nonzero and orthogonal to each other. Let $\mathbf{z}$ be a critical point of f. Then $\mathbf{z}$ can be expanded as

$$\mathbf{z}=\sum_{j=1}^{n}c_j\mathbf{x}_j\quad\text{ with }\quad c_j=\frac{\mathbf{x}_j^T\mathbf{z}}{\|\mathbf{x}_j\|^2},$$

and direct calculation gives

$$\mathbf{0}=\nabla f(\mathbf{z})=\|\mathbf{z}\|^4\mathbf{z}-\sum_{j=1}^{r}(\mathbf{x}_j^T\mathbf{z})^2\mathbf{x}_j=\sum_{j=1}^{r}\left(c_j\|\mathbf{z}\|^4-c_j^2\|\mathbf{x}_j\|^4\right)\mathbf{x}_j+\sum_{j=r+1}^{n}c_j\|\mathbf{z}\|^4\mathbf{x}_j.$$

Therefore,

$$c_j=\begin{cases}0 \quad \text{or} \quad \dfrac{\|\mathbf{z}\|^4}{\|\mathbf{x}_j\|^4}, & j=1,\cdots,r,\\ 0, & j=r+1,\cdots,n.\end{cases}$$

Thus

$$\mathbf{z}=\sum_{j\in\mathcal{I}}\frac{\|\mathbf{z}\|^4}{\|\mathbf{x}_j\|^4}\mathbf{x}_j, \qquad \mathcal{I}\subseteq\{1,\cdots,r\}. \tag{2.1}$$

Next, we characterize the critical points. Direct calculation gives the Hessian matrix of f at any point $\mathbf{y}$

$$\nabla^2 f(\mathbf{y})=\|\mathbf{y}\|^4 I+4\|\mathbf{y}\|^2\mathbf{y}\mathbf{y}^T-2\sum_{i=1}^{r}(\mathbf{x}_i^T\mathbf{y})\mathbf{x}_i\mathbf{x}_i^T. \tag{2.2}$$

1. When $|\mathcal{I}|=1$, the critical points are $\mathbf{z}=\mathbf{x}_j$, $j=1,\cdots,r$. In this case,

   $$\nabla^2 f(\mathbf{x}_j)=\|\mathbf{x}_j\|^4 I+2\|\mathbf{x}_j\|^2\mathbf{x}_j\mathbf{x}_j^T\succ 0.$$

   Therefore, they are local minimizers. As we will see in the rest of the proof, other critical points are not minimizers. Thus, $\mathbf{z}=\mathbf{x}_j$, $j=1,\cdots,r$ are all local minimizers.

2. When a critical point $\mathbf{z}$ satisfies $|\mathcal{I}|\geq 2$, the subspaces $\{\mathbf{d}:\mathbf{z}^T\mathbf{d}=0\}$ and $\text{span}\{\mathbf{x}_j:j\in\mathcal{I}\}$ are $(n-1)$-dimensional and at least 2-dimensional respectively. Therefore, their intersection must be a nontrivial subspace. Choose $\mathbf{d}$ be a nonzero vector in the intersection. That is, $\mathbf{d}\in\{\mathbf{d}:\mathbf{z}^T\mathbf{d}=0\}\cap\text{span}\{\mathbf{x}_j:j\in\mathcal{I}\}$. Then (2.1) and (2.2) give

   $$\begin{aligned}\langle\mathbf{d},\nabla^2 f(\mathbf{z})\mathbf{d}\rangle&=\|\mathbf{z}\|^4\|\mathbf{d}\|^2+4\|\mathbf{z}\|^2(\mathbf{z}^T\mathbf{d})^2-2\sum_{j\in\mathcal{I}}(\mathbf{d}^T\mathbf{x}_j)^2\frac{\|\mathbf{z}\|^4}{\|\mathbf{x}_j\|^2}\\&=\|\mathbf{z}\|^4\|\mathbf{d}\|^2-2\|\mathbf{z}\|^4\|\mathbf{d}\|^2<0.\end{aligned}$$

   This implies the critical points with $|\mathcal{I}|\geq 2$ are strict saddle points.

3. Finally we consider $\mathbf{z}=\mathbf{0}$. Obviously, there exists $\mathbf{u}_0$ such that $\langle\mathcal{A},\mathbf{u}_0^{\otimes 3}\rangle\neq 0$. Define

$$g(\alpha)=f(\alpha\mathbf{u}_0)=\frac{1}{6}\|\mathcal{A}-\alpha^3\mathbf{u}_0^{\otimes 3}\|^2=\frac{1}{6}(\alpha^6\|\mathbf{u}_0\|^6-2\alpha^3\langle\mathcal{A},\mathbf{u}_0^{\otimes 3}\rangle+\|\mathcal{A}\|_F^2).$$

By calculation, we have

$$g'(\alpha)=\alpha^5\|\mathbf{u}_0\|^6-\alpha^2\langle\mathcal{A},\mathbf{u}_0^{\otimes 3}\rangle$$

for which the second term dominant the first term when α is sufficiently close to 0. Therefore, $g'(\alpha)$ does not change the sign in a small neighborhood of $\alpha=0$, and thus $g(\alpha)$ is monotonic near $\alpha=0$. Since $g(\alpha)\neq g(0)$ if $\alpha\neq 0$ near $\alpha=0$, $\mathbf{z}=\mathbf{0}$ is not a local minimum.

This completes the proof. □

Furthermore, the following theorem reveals that f is strongly convex and has a bounded Hessian in a neighbourhood of the local minimizers.

Theorem 2.2 (Strong convexity and Smoothness around Local Minimizers)**.** *Let*

$$\mathcal{A}=\sum_{i=1}^{r}\mathbf{x}_i\otimes\mathbf{x}_i\otimes\mathbf{x}_i,$$

where $\{\mathbf{x}_i\}_{i=1}^r$ *are nonzero orthogonal vectors and* $\|\mathbf{x}_1\|\geq\|\mathbf{x}_2\|\geq\cdots\geq\|\mathbf{x}_r\|>0$. *Let*

$$\kappa=\frac{\|\mathbf{x}_1\|^3}{\|\mathbf{x}_r\|^3}$$

be a condition number of $\mathcal{A}$. *Then, for any positive number* γ *satisfying* $\gamma\leq\min\{0.01,0.15\kappa^{-1}\}$, *any* $i\in\{1,\cdots,r\}$, *and any* $\mathbf{z}\in\mathcal{S}_{i,\gamma}:=\{\mathbf{z}|\|\mathbf{z}-\mathbf{x}_i\|\leq\gamma\|\mathbf{x}_i\|\}$, *the Hessian matrix of*

$$f=\frac{1}{6}\|\mathcal{A}-\mathbf{z}^{\otimes 3}\|_F^2$$

obeys

$$\omega\|\mathbf{x}_i\|^4 I\preceq\nabla^2 f(\mathbf{z})\preceq\zeta\|\mathbf{x}_i\|^4 I,$$

where ω *and* ζ *are absolute positive constants satisfying* $0.05\leq\omega<\zeta\leq 4.1$.

Proof. Let $i\in[r]$ be a fixed index, and $\mathbf{z}\in\mathcal{S}_{i,\gamma}$. Then

$$(1-\gamma)\|\mathbf{x}_i\|\leq\|\mathbf{z}\|\leq(1+\gamma)\|\mathbf{x}_i\|\quad\text{and}\quad\sum_{j\neq i}\frac{|\mathbf{x}_j^T\mathbf{z}|^2}{\|\mathbf{x}_j\|^2}+\left(\|\mathbf{x}_i\|-\frac{\mathbf{x}_i^T\mathbf{z}}{\|\mathbf{x}_i\|}\right)^2\leq\gamma^2\|\mathbf{x}_i\|^2,\ (2.3)$$

which derives

$$(1-\gamma)\|\mathbf{x}_i\| \le \frac{\mathbf{x}_i^T\mathbf{z}}{\|\mathbf{x}_i\|} \le (1+\gamma)\|\mathbf{x}_i\|.$$

Then, (2.2) gives

$$\begin{aligned}\nabla^2 f(\mathbf{z}) &= \|\mathbf{z}\|^4 I + 4\|\mathbf{z}\|^2\mathbf{z}\mathbf{z}^T - 2\sum_{j=1}^{r}(\mathbf{x}_j^T\mathbf{z})\mathbf{x}_j\mathbf{x}_j^T \\ &= \|\mathbf{z}\|^4 I + \underbrace{2(2\|\mathbf{z}\|^2 - \mathbf{x}_i^T\mathbf{z})\mathbf{z}\mathbf{z}^T}_{M_1} + \underbrace{2\mathbf{x}_i^T\mathbf{z}(\mathbf{z}\mathbf{z}^T - \mathbf{x}_i\mathbf{x}_i^T)}_{M_2} - \underbrace{2\sum_{j\neq i}(\mathbf{x}_j^T\mathbf{z})\mathbf{x}_j\mathbf{x}_j^T}_{M_3}.\end{aligned}$$

Let us estimate the terms in the Hessian.

- For M_1: (2.3) implies

  $$\begin{aligned}2\|\mathbf{z}\|^2 - \mathbf{x}_i^T\mathbf{z} &\ge 2(1-\gamma)^2\|\mathbf{x}_i\|^2 - (1+\gamma)\|\mathbf{x}_i\|^2 = (1-5\gamma+2\gamma^2)\|\mathbf{x}_i\|^2, \\ 2\|\mathbf{z}\|^2 - \mathbf{x}_i^T\mathbf{z} &\le 2(1+\gamma)^2\|\mathbf{x}_i\|^2 - (1-\gamma)\|\mathbf{x}_i\|^2 = (1+5\gamma+2\gamma^2)\|\mathbf{x}_i\|^2.\end{aligned}$$

  Thus, provided $0<\gamma\le\frac{1}{5}$, we have $1-5\gamma+2\gamma^2 \ge 1-5\gamma \ge 0$. Therefore

  $$0 \preceq M_1 \preceq (1+5\gamma+2\gamma^2)\|\mathbf{x}_i\|^2\mathbf{z}\mathbf{z}^T,$$

  which yields

  $$\|M_1\| \le (1+5\gamma+2\gamma^2)\|\mathbf{x}_i\|^2\|\mathbf{z}\|^2 \le (1+5\gamma+2\gamma^2)(1+\gamma)^2\|\mathbf{x}_i\|^4.$$

- For M_2: since

  $$\|\mathbf{z}\mathbf{z}^T - \mathbf{x}_i\mathbf{x}_i^T\|_F^2 = \|\mathbf{z}\|^4 + \|\mathbf{x}_i\|^4 - 2(\mathbf{x}_i^T\mathbf{z})^2,$$

  it is deducted from (2.3) that

  $$\begin{aligned}\|\mathbf{z}\mathbf{z}^T - \mathbf{x}_i\mathbf{x}_i^T\| &\le \|\mathbf{z}\mathbf{z}^T - \mathbf{x}_i\mathbf{x}_i^T\|_F \le \left((1+\gamma)^4 + 1 - 2(1-\gamma)^2\right)^{1/2}\|\mathbf{x}_i\|^2 \\ &= \gamma^{1/2}(8+4\gamma+4\gamma^2+\gamma^3)^{1/2}\|\mathbf{x}_i\|^2 \le 3\gamma^{1/2}\|\mathbf{x}_i\|^2,\end{aligned}$$

  if $0\le\gamma\le 1/5$, and thus

  $$\|M_2\| \le |\mathbf{x}_i^T\mathbf{z}|\|\mathbf{z}\mathbf{z}^T - \mathbf{x}_i\mathbf{x}_i^T\|_F \le 3(1+\gamma)\gamma^{1/2}\|\mathbf{x}_i\|^4.$$

- For M_3: since $\{\mathbf{x}_j\}_{j\neq i}$ are orthogonal, (2.3) leads to

  $$\begin{aligned}\|M_3\| &\le \max_{j\neq i}|\mathbf{x}_j^T\mathbf{z}|\|\mathbf{x}_j\|^2 \le \max_{j\neq i}\frac{|\mathbf{x}_j^T\mathbf{z}|}{\|\mathbf{x}_j\|}\cdot\max_{j\neq i}\|\mathbf{x}_j\|^3 \\ &\le \left(\sum_{j\neq i}\frac{|\mathbf{x}_j^T\mathbf{z}|^2}{\|\mathbf{x}_j\|^2}\right)^{1/2}\cdot\max_{j\neq i}\|\mathbf{x}_j\|^3 \le \kappa\gamma\|\mathbf{x}_i\|^4.\end{aligned}$$

Altogether, it follows from Weyl's theorem that, if $\gamma\leq\frac{1}{5}$,

$$\begin{aligned}\lambda_{\min}(\nabla^2 f(\mathbf{z}))\geq&\|\mathbf{z}\|^4-2\|M_2\|-2\|M_3\|\\ \geq&\|\mathbf{z}\|^4-6(1+\gamma)\gamma^{1/2}\|\mathbf{x}_i\|^4-2\kappa\gamma\|\mathbf{x}_i\|^4\\ \geq&\underbrace{\big((1-\gamma)^4-6(1+\gamma)\gamma^{1/2}-2\kappa\gamma\big)}_{\omega}\|\mathbf{x}_i\|^4\end{aligned}$$

and

$$\begin{aligned}\lambda_{\max}(\nabla^2 f(\mathbf{z}))\leq&\|\mathbf{z}\|^4+2\|M_1\|+2\|M_2\|+2\|M_3\|\\ \leq&\|\mathbf{z}\|^4+2(1+5\gamma+2\gamma^2)(1+\gamma)^2\|\mathbf{x}_i\|^4+6(1+\gamma)\gamma^{1/2}\|\mathbf{x}_i\|^4+2\kappa\gamma\|\mathbf{x}_i\|^4\\ \leq&\underbrace{\big((1+\gamma)^4+2(1+\gamma)^2(1+5\gamma+2\gamma^2)+6(1+\gamma)\gamma^{1/2}+2\kappa\gamma\big)}_{\zeta}\|\mathbf{x}_i\|^4.\end{aligned}$$

If we further restrict $0\leq\gamma\leq 0.01$, we have

$$\omega\geq 0.99^4-6\cdot 1.01\cdot 0.1-2\gamma\kappa\geq 0.35-2\gamma\kappa.$$

Therefore, if $0\leq\gamma\leq\min\{0.01,0.15\kappa^{-1}\}$, then $\omega\geq 0.05$. Similarly, $\zeta\leq 4.1$. □

3 Gradient descent algorithm finds the best rank-1 approximation

Model (1.4) is obviously nonconvex, which causes computational challenges. Nowadays, gradient descent algorithm shows good performance for nonconvex problem to solve phase retrieval, matrix completion and blind deconvolution [3, 5]. We may as well apply gradient descent to solve model (1.4) to find the best rank-1 approximation of a given symmetric tensor. The gradient descent scheme is

$$\mathbf{z}_{k+1}=\mathbf{z}_k-\mu_k\nabla f(\mathbf{z}_k),\tag{3.1}$$

with

$$\nabla f(\mathbf{z})=\|\mathbf{z}\|^4\mathbf{z}-\mathcal{A}\otimes_2\mathbf{z}\otimes_3\mathbf{z}$$

and μ_k is the stepsize in the k-th iteration, where the k-mode product of a tensor $\mathcal{A}\in\bigotimes^m\mathbb{R}^n$ with a matrix $U\in\mathbb{R}^{J\times n}$ is denoted by $\mathcal{A}\otimes_k U$ and defined as

$$[\mathcal{A}\otimes_k U]_{i_1,\cdots,i_{k-1},j,i_{k+1},\cdots,i_m}=\sum_{i_k=1}^{n}\mathcal{A}_{i_1,\cdots,i_m}U_{j,i_k}.$$

As we have seen from Theorem 2.1, the objective function in (1.4) has different types of critical points, namely, local minima, strict saddle points, and **0**. Though gradient descent algorithm is shown not to converge to strict saddle points [17] with probability 1, it might converges to the critical point **0** or a local minimum. Furthermore, the convergence rate is unknown. Therefore, there is a need of the analysis of the convergence of gradient descent for (1.4).

In order the algorithm converges to the global minimizer, the initial guess should lie in a neighbourhood of the global minimizer in general. We design a careful initialization, which is generated as follows. Let $\mathbf{w}_i \in \mathbb{R}^n$, $i=1,\cdots,L$ be i.i.d. random vectors chosen uniformly from sphere with radius $\frac{1}{\sqrt{n}}$ and

$$\mathbf{z}_0 = \frac{1}{L}\sum_{i=1}^{L}(\mathbf{w}_i - n^2\nabla f(\mathbf{w}_i)). \tag{3.2}$$

In other words, we average outputs of the first step of randomly initialized gradient descent algorithms as the initial guess.

We will prove that gradient descent, starting from the initial guess in (3.2), converges to a global minimizer, and the convergence rate is linear. The results are summarized in the following theorem.

Theorem 3.1 (Convergence to the Global Minimizer)**.** *Let $\mathcal{A}$, $\mathbf{x}_i$, $i=1,\cdots,r$, and κ be the same as in Theorem* 2.2*. Let*

$$\delta = \frac{\|\mathbf{x}_1\|^6 - \|\mathbf{x}_2\|^6}{3\|\mathbf{x}_1\|^6}$$

be a gap. Assume $\delta > 0$. Let $\{\mathbf{z}_k\}_{k\in\mathbb{N}}$ be generated by (3.1) *with the initial guess $\mathbf{z}_0$ in* (3.2) *and the step size μ_k satisfying $1-\mu_k\|\mathbf{z}_k\|^4 = \theta$ for all $k \geq 0$. Let γ be a positive number satisfying $0 < \gamma \leq \min\{0.01, 0.15\kappa^{-1}\}$. There exist positive constants C_δ, c_δ, $C_{\theta,\delta,\|\mathbf{x}_1\|}$, $C_{\theta,\|\mathbf{x}_1\|}$, $\theta_\gamma \in (0,1)$, $\rho_\theta \in (0,1)$ depending only on their corresponding subscripts respectively such that: with probability at least $1-e^{-c_\delta n}$, as long as $L \geq C_\delta n$ and $\theta \in (\theta_\gamma, 1)$, it holds true that*

$$\|\mathbf{z}_k - \mathbf{x}_1\| \leq \gamma \cdot \rho_\theta^{k-T_\gamma}\|\mathbf{x}_1\|, \qquad \forall k \geq T_\gamma,$$

where

$$T_\gamma = C_{\theta,\delta} r^{2.1}\kappa^{4.2}(\log\gamma^{-1} + \log r + \log\kappa) + C_{\theta,\|\mathbf{x}_1\|}\gamma^{-1}(\log r + \log\kappa).$$

The above theorem tells us that, after T_γ steps, the gradient descent algorithm converges linearly to the principal component in the CP decomposition of $\mathcal{A}$. To obtain a CP decomposition of $\mathcal{A}$, we may combine the randomly initialized gradient

descent algorithm with the greedy strategy. The combined approach consists of r phases. At phase j, we first apply the randomly initialized gradient descent algorithm to obtain $\mathbf{x}_{i_j}$, and then we subtract $\mathbf{x}_{i_j}^{\otimes 3}$ from $\mathcal{A}$. Due to the orthogonality of factors of $\mathcal{A}$, it gives another new factor of $\mathcal{A}$ in the next phase. By this way, after r phases, we obtain all the factors of $\mathcal{A}$, and hence the CP decomposition of $\mathcal{A}$.

A theoretical result concerning the convergence of power method for symmetric rank-1 approximation of orthogonally decomposible tensors was provided in [1]. Although the result there shows the super-linear convergence, it is not guaranteed the convergence to the global minimum, and only the convergence to a local minimum was established. As a comparison, our result indicates the gradient descent algorithm converges with high probability to the global minimum.

To prove Theorem 3.1, we divide the iteration into two stages, and we give convergence analysis for these two stages respectively. The first stage is the initial finite steps of the iteration. We show that, after

$$C_{\theta,\delta}r^{2.1}\kappa^{4.2}(\log\gamma^{-1}+\log r+\log\kappa)+C_{\theta,\|\mathbf{x}_1\|}\gamma^{-1}(\log r+\log\kappa)$$

steps, $\mathbf{z}_k$ will be in a γ-neighbourhood of the globle minimizer $\mathbf{x}_1$ with high probability. The result is presented in the following Theorem 3.2.

Theorem 3.2 (Initial Stage)**.** *Let $\mathcal{A}$, κ, δ, $\{\mathbf{x}_i\}_{i=1}^r$, $\{\mathbf{z}_k,\mu_k\}_{k\in\mathbb{N}}$ be the same as in Theorem* 3.1*. Let $\theta\in[0.7,1)$ and $\gamma\in(0,0.05]$. There exist positive constants C_δ, c_δ, $C_{\theta,\delta,\|\mathbf{x}_1\|}$, $C_{\theta,\|\mathbf{x}_1\|}$ depending only on their corresponding subscripts respectively such that: with probability at least $1-e^{-c_\delta n}$, as long as $L\geq C_\delta n$, it holds true that*

$$\|\mathbf{z}_k-\mathbf{x}_1\|\leq\gamma\|\mathbf{x}_1\|$$

for $k=T$, where

$$T\leq C_{\theta,\delta}r^{2.1}\kappa^{4.2}(\log\gamma^{-1}+\log r+\log\kappa)+C_{\theta,\|\mathbf{x}_1\|}\gamma^{-1}(\log r+\log\kappa).$$

Proof. The proof is delayed in Section 4. □

In the second stage, the iterations are kept in the neighbourhood $\mathcal{S}_\gamma$ of the global minimizer $\mathbf{x}_1$. Since f is strongly convex in $\mathcal{S}_\gamma$ according to Theorem 2.2, the gradient descent algorithm converges linearly to $\mathbf{x}_1$. In particular, we have the following theorem.

Theorem 3.3 (Refinement Stage)**.** *Let $\mathcal{A}$, κ, δ, θ, $\{\mathbf{x}_i\}_{i=1}^r$, $\{\mathbf{z}_k,\mu_k\}_{k\in\mathbb{N}}$ be the same as in Theorem* 3.1*. Let γ be satisfying $0\leq\gamma\leq\min\{0.01,0.15\kappa^{-1}\}$. There exists positive constants $\theta_c\in(0,1)$ and $\rho_\theta\in[0,1)$ such that: if $\theta\in(\theta_c,1)$ and $\mathbf{z}_T\in\mathcal{S}_\gamma:=\{\mathbf{z}:\|\mathbf{z}-\mathbf{x}_1\|\leq\gamma\|\mathbf{x}_1\|\}$ for some T, then $\{\mathbf{z}_k\}_{k\geq T}\subset\mathcal{S}_\gamma$ and*

$$\|\mathbf{z}_{k+1}-\mathbf{x}_1\|\leq\rho_\theta\|\mathbf{z}_k-\mathbf{x}_1\|,\qquad\forall k\geq T.$$

Proof. It suffices to prove

$$\|\mathbf{z}_{k+1}-\mathbf{x}_1\|\leq(1-\rho_\theta)\|\mathbf{z}_k-\mathbf{x}_1\|,\qquad \forall k\geq T,$$

which is done by induction. Let k be larger than T. Suppose $\mathbf{z}_k\in\mathcal{S}_\gamma$. From Theorem 2.2,

$$\omega\|\mathbf{x}_1\|^4\leq\lambda_{\min}(\nabla^2 f(\mathbf{z}))\leq\lambda_{\max}(\nabla^2 f(\mathbf{z}))\leq\zeta\|\mathbf{x}_1\|^4,\qquad \forall\mathbf{z}\in\mathcal{S}_\gamma,$$

from which it follows that, for any $\mathbf{z},\mathbf{y}\in\mathcal{S}_\gamma$,

$$(\mathbf{z}-\mathbf{y})^T(\nabla f(\mathbf{z})-\nabla f(\mathbf{y}))\geq\omega\|\mathbf{x}_1\|^4\|\mathbf{z}-\mathbf{y}\|^2,$$

and

$$\|\nabla f(\mathbf{z})-\nabla f(\mathbf{y})\|\leq\zeta\|\mathbf{x}_1\|^4\|\mathbf{z}-\mathbf{y}\|.$$

In view of gradient descent scheme, the above two inequalities imply

$$\begin{aligned}\|\mathbf{z}_{k+1}-\mathbf{x}_1\|^2&=\|(\mathbf{z}_k-\mu_k\nabla f(\mathbf{z}_k))-(\mathbf{x}_1-\mu_k\nabla f(\mathbf{x}_1))\|^2\\&=\|\mathbf{z}_k-\mathbf{x}_1\|^2-2\mu_k(\mathbf{z}_k-\mathbf{x}_1)^T(\nabla f(\mathbf{z}_k)-\nabla f(\mathbf{x}_1))+\mu_k^2\|\nabla f(\mathbf{z}_k)-\nabla f(\mathbf{x}_1)\|^2\\&\leq\left(1-2\mu_k\omega\|\mathbf{x}_1\|^4+\mu_k^2\zeta^2\|\mathbf{x}_1\|^8\right)\|\mathbf{z}_k-\mathbf{x}_1\|^2\\&:=\rho^2\|\mathbf{z}_k-\mathbf{x}_1\|^2.\end{aligned}$$

Since $\theta=1-\mu_k\|\mathbf{z}_k\|^4$ and $\mathbf{z}_k\in\mathcal{S}_\gamma$ with $\gamma\in(0,\min\{0.01,0.15\kappa^{-1}\}]$, we have

$$\frac{1-\theta}{1.01^4}\frac{1}{\|\mathbf{x}_1\|^4}\leq\frac{1-\theta}{(1+\gamma)^4}\frac{1}{\|\mathbf{x}_1\|^4}\leq\mu_k=\frac{1-\theta}{\|\mathbf{z}_k\|^4}\leq\frac{1-\theta}{(1-\gamma)^4}\frac{1}{\|\mathbf{x}_1\|^4}\leq\frac{1-\theta}{0.99^4}\frac{1}{\|\mathbf{x}_1\|^4}.$$

Therefore, if we choose θ_c such that $\frac{1-\theta_c}{0.99^4}=\frac{\omega}{\zeta^2}$, then for any $\theta\in(\theta_c,1)$, we have

$$\frac{1-\theta}{1.01^4}\leq\mu_k\|\mathbf{x}_1\|^4\leq\frac{1-\theta}{0.99^4}\leq\frac{\omega}{\zeta^2},$$

which implies

$$0\leq\rho^2=1-\mu_k\|\mathbf{x}_1\|^4\left(2\omega-\mu_k\zeta^2\|\mathbf{x}_1\|^4\right)\leq1-\frac{1-\theta}{1.01^4}\omega:=\rho_\theta^2<1.$$

Thus, we complete the proof. □

Combining Theorems 3.2 and 3.3, we obtain Theorem 3.1.

4 Proof of Theorem 3.2

This section is devoted to the proof of Theorem 3.2.

Proof of Theorem 3.2. We can expand $\mathbf{x}_i$, $i=1,\cdots,r$ to a basis of $\mathbb{R}^n$, denoted by $\{\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_n\}$, where $\mathbf{u}_i=\mathbf{x}_i/\|\mathbf{x}_i\|$, $i=1,\cdots,r$. Set $\alpha_k^{(i)}=\mathbf{z}_k^T\mathbf{u}_i$, $i=1,\cdots,r$ and

$$\beta_k=\Big(\sum_{j=r+1}^{n}(\mathbf{u}_j^T\mathbf{z}_k)^2\Big)^{1/2}.$$

Thus, we have

$$\|\mathbf{z}_k\|^2=\sum_{i=1}^{r}\big(\alpha_k^{(i)}\big)^2+\beta_k^2,$$

and the gradient descent scheme $\mathbf{z}_{k+1}=\mathbf{z}_k-\mu_k\nabla f(\mathbf{z}_k)$ can be expressed as

$$\begin{cases}\alpha_{k+1}^{(1)}=\alpha_k^{(1)}-\mu_k(\|\mathbf{z}_k\|^4\alpha_k^{(1)}-\|\mathbf{x}_1\|^3(\alpha_k^{(1)})^2),\\ \quad\vdots\\ \alpha_{k+1}^{(r)}=\alpha_k^{(r)}-\mu_k(\|\mathbf{z}_k\|^4\alpha_k^{(r)}-\|\mathbf{x}_r\|^3(\alpha_k^{(r)})^2),\\ \beta_{k+1}=\big|1-\mu_k\|\mathbf{z}_k\|^4\big|\beta_k.\end{cases}\tag{4.1}$$

Since $\mathbf{w}_i\in\mathbb{R}^n$, $i=1,\cdots,L$ are i.i.d. random vectors chosen uniformly from sphere with radius $\frac{1}{\sqrt{n}}$ and set $\mu_{-1}=n^2$, we have

$$\begin{aligned}\mathbf{z}_0&=\frac{1}{L}\sum_{i=1}^{L}(\mathbf{w}_i-\mu_{-1}\nabla f(\mathbf{w}_i))\\&=\frac{1}{L}\sum_{i=1}^{L}(1-\mu_{-1}\|\mathbf{w}_i\|^4)\mathbf{w}_i+\mu_{-1}\frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{r}(\mathbf{x}_j^T\mathbf{w}_i)^2\mathbf{x}_j\\&=\frac{n^2}{L}\sum_{i=1}^{L}\sum_{j=1}^{r}(\mathbf{x}_j^T\mathbf{w}_i)^2\mathbf{x}_j,\end{aligned}$$

and

$$\alpha_0^{(l)}=\mathbf{z}_0^T\mathbf{u}_l=\frac{\mathbf{z}_0^T\mathbf{x}_l}{\|\mathbf{x}_l\|}=\frac{n^2}{L}\sum_{i=1}^{L}(\mathbf{x}_l^T\mathbf{w}_i)^2\|\mathbf{x}_l\|,\qquad l=1,2,\cdots,r,\tag{4.2a}$$

$$\beta_0=\Big(\sum_{j=r+1}^{n}(\mathbf{u}_j^T\mathbf{z}_0)^2\Big)^{1/2}=0.\tag{4.2b}$$

The rest of the proof is divided into the following several steps.

Step 1. We show the following statement. For any $\delta>0$, there exist constants $C_\delta,c_\delta>0$ such that: when $L>C_\delta n$, we have

$$(1+\delta)\|\mathbf{x}_i\|^3\geq\alpha_0^{(i)}\geq(1-\delta)\|\mathbf{x}_i\|^3,\qquad i=1,\cdots,r, \tag{4.3}$$

holds with probability at least $1-2e^{-c_\delta n}$. In particular, we choose

$$\delta=\frac{\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6}{3\|\mathbf{x}_1\|^6}\in\Big(0,\frac{1}{3}\Big),$$

we have

$$\frac{4}{3}\|\mathbf{x}_i\|^3\geq\alpha_0^{(i)}\geq\frac{2}{3}\|\mathbf{x}_i\|^3,\qquad i=1,\cdots,r, \tag{4.4a}$$

$$\|\mathbf{x}_1\|^3\alpha_0^{(1)}-\|\mathbf{x}_i\|^3\alpha_0^{(i)}\geq\frac{\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6}{3},\qquad\forall i>1, \tag{4.4b}$$

with probability at least $1-2e^{-c_\delta n}$. Since

$$\mathbb{E}\Big(\frac{n^2}{L}\sum_{\ell=1}^{L}\mathbf{w}_\ell\mathbf{w}_\ell^T\Big)=I,$$

by [26, Remark 5.40], for any positive constant $\delta>0$, there exist constants $C_\delta,c_\delta>0$ such that: when $L\geq C_\delta n$, we get

$$\left\|\frac{n^2}{L}\sum_{\ell=1}^{L}\mathbf{w}_\ell\mathbf{w}_\ell^T-I\right\|\leq\delta \tag{4.5}$$

with probability at most $1-2e^{-c_\delta n}$. Applying it to $\alpha_0^{(i)}$ in (4.2), we obtain (4.3).

Moreover, (4.5) implies, for $\ell\neq1$,

$$\begin{aligned}
&\|\mathbf{x}_1\|^3\alpha_0^{(1)}-\|\mathbf{x}_i\|^3\alpha_0^{(i)}-\|\mathbf{x}_1\|^6+\|\mathbf{x}_i\|^6\\
=&\|\mathbf{x}_1\|^4\mathbf{x}_1^T\left(\frac{n^2}{L}\sum_{\ell=1}^{L}\mathbf{w}_\ell\mathbf{w}_\ell^T-I\right)\mathbf{x}_1-\|\mathbf{x}_i\|^4\mathbf{x}_i^T\left(\frac{n^2}{L}\sum_{\ell=1}^{L}\mathbf{w}_\ell\mathbf{w}_\ell^T-I\right)\mathbf{x}_i\\
\geq&-\delta\|\mathbf{x}_1\|^6-\delta\|\mathbf{x}_i\|^6.
\end{aligned}$$

Therefore,

$$\|\mathbf{x}_1\|^3\alpha_0^{(1)}-\|\mathbf{x}_i\|^3\alpha_0^{(i)}\geq(1-\delta)\|\mathbf{x}_1\|^6-(1+\delta)\|\mathbf{x}_i\|^6,$$

which by choosing

$$\delta=\frac{\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6}{3\|\mathbf{x}_1\|^6}\leq\frac{\|\mathbf{x}_1\|^6-\|\mathbf{x}_\ell\|^6}{3\|\mathbf{x}_1\|^6}$$

gives (4.4b).

The rest of the proof are established on the event, where (4.3), (4.4a) and (4.4b) are successful, and the probability is at least $1-2e^{-c_\delta n}$.

Step 2. Define

$$c_k=\min_{i>1}\{\|\mathbf{x}_1\|^3-\|\mathbf{x}_i\|^3\alpha_k^{(i)}/\alpha_k^{(1)}\}.$$

We prove

$$c_k\geq\frac{\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6}{4\|\mathbf{x}_1\|^3},\qquad\forall k\geq 0. \tag{4.6}$$

To this end, we first show by induction that

$$\|\mathbf{x}_1\|^3\alpha_k^{(1)}>\|\mathbf{x}_i\|^3\alpha_k^{(i)},\qquad\forall k\geq 0,\qquad 1<i\leq r, \tag{4.7}$$

which holds obviously for $k=0$. Suppose (4.7) holds for $k=0,\cdots,m$, which yields

$$\theta+\mu_m\|\mathbf{x}_1\|^3\alpha_m^{(1)}\geq\theta+\mu_m\|\mathbf{x}_i\|^3\alpha_m^{(i)},\qquad\forall i>1.$$

Applying this inequality to (4.1) gives

$$\frac{\alpha_{m+1}^{(1)}}{\alpha_{m+1}^{(i)}}=\frac{\theta+\mu_m\|\mathbf{x}_1\|^3\alpha_m^{(1)}}{\theta+\mu_m\|\mathbf{x}_i\|^3\alpha_m^{(i)}}\cdot\frac{\alpha_m^{(1)}}{\alpha_m^{(i)}}\geq\frac{\alpha_m^{(1)}}{\alpha_m^{(i)}}\geq\cdots\geq\frac{\alpha_0^{(1)}}{\alpha_0^{(i)}}>\frac{\|\mathbf{x}_i\|^3}{\|\mathbf{x}_1\|^3},\qquad\forall i>1, \tag{4.8}$$

which implies (4.7) with $k=m+1$ immediately. Therefore, (4.7) holds for any $k\geq 0$. It is also seen from (4.8) that the ratio $\frac{\alpha_k^{(1)}}{\alpha_k^{(i)}}$ is monotonically increasing with respect to k, and hence c_k is. Thus, $c_k\geq c_0$. By (4.4b) in Step 1, we have

$$c_0\geq\frac{\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6}{4\|\mathbf{x}_1\|^3},$$

we obtain (4.6).

Step 3. We then prove the following result. For any $\gamma>0$, we have

$$\rho_k:=\frac{\alpha_k^{(1)}}{\sqrt{\sum_{i=2}^r(\alpha_k^{(i)})^2+\beta_k^2}}\geq\frac{2+\gamma}{\gamma}$$

for all $k\geq T_1$ with

$$T_1\leq C_{\theta,\|\mathbf{x}_1\|}(\log r+\log\kappa)+C_{\theta,\delta}\cdot r^{2.1}\kappa^{4.2}\left(\log\gamma^{-1}+\log r+\log\kappa\right)$$

for some constants $C_{\theta,\|\mathbf{x}_1\|}$ and $C_{\theta,\delta}$.

We first give an obvious lower bound of ρ_k. From (4.1) and (4.2) we see that $\beta_0=0$ and $\beta_{k+1}=\theta\beta_k$, which implies $\beta_k=0$ for any $k\geq 0$. Recall

$$\kappa=\max_i \frac{\|\mathbf{x}_1\|^3}{\|\mathbf{x}_i\|^3}$$

is a condition number of $\mathcal{A}$. A lower bound of ρ_k is obtained as in the following

$$\rho_k=\frac{\alpha_k^{(1)}}{\sqrt{\sum_{i=2}^r(\alpha_k^{(i)})^2+\beta_k^2}}=\frac{\alpha_k^{(1)}}{\sqrt{\sum_{i=2}^r(\alpha_k^{(i)})^2}}\geq\frac{1}{\sqrt{r\kappa^2}}, \tag{4.9}$$

where in the last inequality we used (4.7).

Now we define

$$\nu_k:=\frac{\alpha_k^{(1)}}{\|\mathbf{x}_1\|}$$

and show that

$$\nu_k\leq\nu:=\max\left\{\theta+0.16\theta^{-2}r^2\kappa^4,\frac{4\|\mathbf{x}_1\|^2}{3}+(1-\theta)\frac{9}{4\|\mathbf{x}_1\|^4}\right\},\qquad\forall k\geq 0, \tag{4.10}$$

whose proof is presented in the rest of this paragraph. By simple calculation,

$$\|\mathbf{z}_k\|^2=(\alpha_k^{(1)})^2+(\alpha_k^{(1)})^2/\rho_k^2=(1+\rho_k^{-2})(\alpha_k^{(1)})^2.$$

It follows from (4.1) that

$$\alpha_{k+1}^{(1)}=\alpha_k^{(1)}-\mu_k\left(\|\mathbf{z}_k\|^4\alpha_k^{(1)}-\|\mathbf{x}_1\|^3(\alpha_k^{(1)})^2\right)=\theta\alpha_k^{(1)}+\frac{(1-\theta)\|\mathbf{x}_1\|^3}{(1+\rho_k^{-2})^2}\frac{1}{(\alpha_k^{(1)})^2},$$

and therefore

$$\nu_{k+1}=\theta\nu_k+(1-\theta)\left(1+\rho_k^{-2}\right)^{-2}\nu_k^{-2}. \tag{4.11}$$

A lower bound of ν_k is obtained by the inequality of arithmetic and geometric means

$$\begin{aligned}\nu_{k+1}&=\frac{\theta}{2}\nu_k+\frac{\theta}{2}\nu_k+(1-\theta)\left(1+\rho_k^{-2}\right)^{-2}\nu_k^{-2}\\&\geq 1.88\theta^{\frac{2}{3}}(1-\theta)^{\frac{1}{3}}\left(1+\rho_{k-1}^{-2}\right)^{-\frac{2}{3}}\\&\geq 1.88\theta^{\frac{2}{3}}(1-\theta)^{\frac{1}{3}}(r\kappa^2)^{-\frac{2}{3}},\end{aligned} \tag{4.12}$$

where we used (4.9). For the upper bound,

$$\nu_{k+1}=\theta\nu_k+(1-\theta)\left(1+\rho_k^{-2}\right)^{-2}\nu_k^{-2}\leq\left(\theta+(1-\theta)\nu_k^{-3}\right)\nu_k. \tag{4.13}$$

Therefore, if $\nu_k>1$, then the sequence $\{\nu_k,\nu_{k+1},\cdots\}$ starts decreasing until $\nu_{k'}\leq 1$ for some $k'>k$. Let $2\leq k_1<k_2<\cdots$ be all the integers satisfying $\nu_{k_i-1}\leq 1<\nu_{k_i}$. Obviously,

$$\sup_{k\geq 0}\nu_k=\sup\{\nu_0,\nu_1,\nu_{k_1},\nu_{k_2},\cdots\}.$$

It is deducted from (4.13) and the lower bound of ν_{k_i-1} that

$$\begin{aligned}\nu_{k_i}&\leq\left(\theta+(1-\theta)\nu_{k_i-1}^{-3}\right)\nu_{k_i-1}\leq\theta+(1-\theta)\left(1.88\theta^{\frac{2}{3}}(1-\theta)^{\frac{1}{3}}\left(1+\rho_{k_i-2}^{-2}\right)^{-\frac{2}{3}}\right)^{-3}\\&\leq\theta+0.16\theta^{-2}\left(1+\rho_{k_i-2}^{-2}\right)^2\leq\theta+0.16\theta^{-2}r^2\kappa^4,\end{aligned}\tag{4.14}$$

where the last inequality follows from (4.9). By (4.4a),

$$\nu_0\leq\frac{4\|\mathbf{x}_1\|^2}{3}.$$

It remains to estimate ν_1. Since

$$\begin{aligned}&\|\mathbf{z}_0\|\geq\alpha_0^{(1)}\geq\frac{2\|\mathbf{x}_1\|^3}{3},\\&\alpha_1^{(1)}=\theta\alpha_0^{(1)}+(1-\theta)\frac{\|\mathbf{x}_1\|^3(\alpha_0^{(1)})^2}{\|\mathbf{z}_0\|^4}\leq\theta\alpha_0^{(1)}+(1-\theta)\frac{\|\mathbf{x}_1\|^3}{\|\mathbf{z}_0\|^2}\leq\frac{4\theta\|\mathbf{x}_1\|^3}{3}+(1-\theta)\frac{9}{4\|\mathbf{x}_1\|^3}.\end{aligned}$$

Therefore,

$$\nu_1\leq\frac{4\theta\|\mathbf{x}_1\|^2}{3}+(1-\theta)\frac{9}{4\|\mathbf{x}_1\|^4}.\tag{4.15}$$

Putting this together with (4.14) we obtain (4.10).

Next we prove in this paragraph

$$\|\mathbf{z}_k\|\leq\iota\|\mathbf{x}_1\|,\qquad\forall k\geq T_{1,1},\tag{4.16}$$

where ι is a constant defined by

$$\iota=\max\left(\theta+(1-\theta)\frac{r^{\frac{8}{15}}\kappa^{\frac{16}{15}}}{5\cdot\left(\frac{1.88}{4}\right)^{\frac{4}{5}}\theta^{\frac{32}{15}}(1-\theta)^{\frac{2}{3}}},1.1\right),\tag{4.17}$$

and $T_{1,1}$ is a constant bounded by

$$T_{1,1}\leq C_{\theta,\|\mathbf{x}_1\|}(\log r+\log\kappa+\log n)$$

for some $C_{\theta,\|\mathbf{x}_1\|}$ depending only on θ and $\|\mathbf{x}_1\|$. For this purpose, we first estimate the lower bound of $\|\mathbf{z}_k\|$. We have for all $k\geq 0$,

$$\begin{aligned}
\|\mathbf{z}_{k+1}\|^2=&\sum_{i=1}^{r}\left(\alpha_{k+1}^{(1)}\right)^2\overset{(4.1)}{\geq}\left(\theta\alpha_k^{(1)}+\frac{1-\theta}{\|\mathbf{z}_k\|^4}\|\mathbf{x}_1\|^3\left(\alpha_k^{(1)}\right)^2\right)^2+\theta^2\sum_{i=2}^{r}\left(\alpha_k^{(i)}\right)^2\\
\geq&\theta^2\|\mathbf{z}_k\|^2+\frac{(1-\theta)^2}{\|\mathbf{z}_k\|^8}\|\mathbf{x}_1\|^6\left(\alpha_k^{(1)}\right)^4\\
\overset{(4.12)}{\geq}&\theta^2\|\mathbf{z}_k\|^2+\frac{\theta^{\frac{8}{3}}(1-\theta)^{\frac{10}{3}}}{\|\mathbf{z}_k\|^8}\|\mathbf{x}_1\|^{10}1.88^4r^{-\frac{8}{3}}\kappa^{-\frac{16}{3}}\\
=&\frac{\theta^2}{4}\|\mathbf{z}_k\|^2+\frac{\theta^2}{4}\|\mathbf{z}_k\|^2+\frac{\theta^2}{4}\|\mathbf{z}_k\|^2+\frac{\theta^2}{4}\|\mathbf{z}_k\|^2+\frac{\theta^{\frac{8}{3}}(1-\theta)^{\frac{10}{3}}}{\|\mathbf{z}_k\|^8}\|\mathbf{x}_1\|^{10}1.88^4r^{-\frac{8}{3}}\kappa^{-\frac{16}{3}}\\
\geq&5\left(\frac{\theta^8}{4^4}\cdot\theta^{\frac{8}{3}}(1-\theta)^{\frac{10}{3}}\|\mathbf{x}_1\|^{10}1.88^4r^{-\frac{8}{3}}\kappa^{-\frac{16}{3}}\right)^{\frac{1}{5}}\\
=&5\cdot\left(\frac{1.88}{4}\right)^{\frac{4}{5}}\theta^{\frac{32}{15}}(1-\theta)^{\frac{2}{3}}r^{-\frac{8}{15}}\kappa^{-\frac{16}{15}}\|\mathbf{x}_1\|^2.
\end{aligned}\tag{4.18}$$

Now we estimate an upper bound of $\|\mathbf{z}_k\|$. The iteration scheme (4.1) and the choice of μ_k give, for all $k\geq 0$,

$$\begin{aligned}
\alpha_{k+1}^{(i)}=&\theta\alpha_k^{(i)}+(1-\theta)\frac{\|\mathbf{x}_i\|^3}{\|\mathbf{z}_k\|^4}(\alpha_k^{(i)})^2\\
\leq&\theta\alpha_k^{(i)}+(1-\theta)\frac{\|\mathbf{x}_i\|^3}{\|\mathbf{z}_k\|^3}\alpha_k^{(i)}\leq\left(\theta+(1-\theta)\frac{\|\mathbf{x}_1\|^3}{\|\mathbf{z}_k\|^3}\right)\alpha_k^{(i)}.
\end{aligned}$$

Summing up the squares of the inequalities for $i=1,\cdots,r$ and noticing $\beta_k=0$ for all k, we obtain

$$\|\mathbf{z}_{k+1}\|\leq\left(\theta+(1-\theta)\frac{\|\mathbf{x}_1\|^3}{\|\mathbf{z}_k\|^3}\right)\|\mathbf{z}_k\|,\qquad\forall k\geq 0.\tag{4.19}$$

Let $T_{1,1}$ be the smallest non-zero such that $\|\mathbf{z}_{T_{1,1}}\|\leq\iota\|\mathbf{x}_1\|$. There are two cases.

- Case 1: $\|\mathbf{z}_{T_{1,1}}\|\geq\|\mathbf{x}_1\|$. In this case, (4.19) gives

$$\|\mathbf{z}_{T_{1,1}+1}\|\leq\|\mathbf{z}_{T_{1,1}}\|\leq\iota\|\mathbf{x}_1\|.$$

- Case 2: $\|\mathbf{z}_{T_{1,1}}\|<\|\mathbf{x}_1\|$. In this case, we have by (4.18) and (4.19)

$$\|\mathbf{z}_{T_{1,1}+1}\|\leq\theta\|\mathbf{z}_{T_{1,1}}\|+(1-\theta)\frac{\|\mathbf{x}_1\|^3}{\|\mathbf{z}_{T_{1,1}}\|^2}\leq\iota\|\mathbf{x}_1\|.$$

By induction of $\|\mathbf{z}_k\|$, we have $\|\mathbf{z}_k\|\leq\iota\|\mathbf{x}_1\|$ for all $k\geq T_{1,1}$. Let us estimate $T_{1,1}$. We have either $T_{1,1}=1$ or $T_{1,1}\geq 2$. The former case has a constant bound, and the latter case is estimated as in the following. The definition of $T_{1,1}$ gives

$$\frac{\|\mathbf{x}_1\|}{\|\mathbf{z}_k\|}<\iota^{-1}$$

for all $k\in(0,T_{1,1})$, which together with (4.19) implies

$$\iota\|\mathbf{x}_1\|<\|\mathbf{z}_{T_{1,1}-1}\|\leq\big(1-(1-\theta)(1-\iota^{-3})\big)^{T_{1,1}-2}\|\mathbf{z}_1\|.$$

Therefore,

$$T_{1,1}\leq\frac{\log\iota+\log\frac{\|\mathbf{x}_1\|}{\|\mathbf{z}_1\|}}{\log(1-(1-\theta)(1-\iota^{-3}))}+2\leq\frac{\log\frac{\|\mathbf{x}_1\|}{\|\mathbf{z}_1\|}}{\log(1-(1-\theta)(1-1.1^{-3}))}+2, \tag{4.20}$$

where the last inequality is deducted from $\iota\geq 1.1$ and the denominator is negative. It remains to estimate $\|\mathbf{z}_1\|$. It follows from (4.7), (4.10), and $\beta_k=0$ that

$$\|\mathbf{z}_k\|^2=\sum_{i=1}^{r}\left(\alpha_k^{(i)}\right)^2\leq r\kappa^2\left(\alpha_k^{(1)}\right)^2\leq r\kappa^2\nu^2\|\mathbf{x}_1\|^2,\qquad\forall k>0, \tag{4.21}$$

where ν is the upper bound of $\nu_k,k\geq 0$ estimated in (4.10). Setting $k=1$ and plugging (4.21) into (4.20), we obtain

$$T_{1,1}\leq C_{\theta,\|\mathbf{x}_1\|}(\log\kappa+\log r),$$

where $C_{\theta,\|\mathbf{x}_1\|}$ is a constant depending only on θ and $\|\mathbf{x}_1\|$.

Finally, we prove in this paragraph that $\rho_k\geq\frac{2+\gamma}{\gamma}$ for all $k\geq T_{1,1}+T_{1,2}$, where

$$T_{1,2}\leq C_{\theta,\delta}r^{2.1}\kappa^{4.2}(\log\gamma^{-1}+\log r+\log\kappa).$$

Set

$$i^*=\arg\max_{i>1}\|\mathbf{x}_i\|^3\alpha_k^{(i)}.$$

Due to $\theta=1-\mu_k\|\mathbf{z}_k\|^4$ and the iteration scheme (4.1), it holds that

$$\sqrt{\sum_{i=2}^{r}(\alpha_{k+1}^{(i)})^2}=\sqrt{\sum_{i=2}^{r}\left((\theta+\mu_k\|\mathbf{x}_i\|^3\alpha_k^{(i)})\alpha_k^{(i)}\right)^2}\leq(\theta+\mu_k\|\mathbf{x}_{i^*}\|^3\alpha_k^{(i^*)})\sqrt{\sum_{i=2}^{r}(\alpha_k^{(i)})^2}.$$

Combining it with the definition of ρ_k and the fact $\beta_k=0$, we obtain

$$\rho_{k+1}\geq\frac{\theta+\mu_k\|\mathbf{x}_1\|^3\alpha_k^{(1)}}{\theta+\mu_k\|\mathbf{x}_{\ell^*}\|^3\alpha_k^{(i^*)}}\rho_k.$$

Using the definition

$$c_k=\min_{i>1}\left\{\|\mathbf{x}_1\|^3-\|\mathbf{x}_i\|^3\alpha_k^{(i)}/\alpha_k^{(1)}\right\},$$

the factor in the above inequality is estimated by

$$\begin{aligned}\frac{\theta+\mu_k\|\mathbf{x}_1\|^3\alpha_k^{(1)}}{\theta+\mu_k\|\mathbf{x}_{i^*}\|^3\alpha_k^{(i^*)}}&=1+\frac{\mu_k(\|\mathbf{x}_1\|^3\alpha_k^{(1)}-\|\mathbf{x}_{i^*}\|^3\alpha^{(i^*)})}{\theta+\mu_k\|\mathbf{x}_{i^*}\|^3\alpha_k^{(i^*)}}\\&\geq 1+\frac{c_k(1-\theta)}{\theta\|\mathbf{z}_k\|^4/\alpha_k^{(1)}+(1-\theta)\|\mathbf{x}_{\ell^*}\|^3\alpha_k^{(i^*)}/\alpha_k^{(1)}}\\&\overset{(4.6)}{\geq} 1+\frac{(\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6)(1-\theta)}{4\|\mathbf{x}_1\|^3(\theta\|\mathbf{z}_k\|^4/\alpha_k^{(1)}+(1-\theta)\|\mathbf{x}_{i^*}\|^3\alpha_k^{(i^*)}/\alpha_k^{(1)})}\\&\overset{(4.16),\ (4.7)}{\geq} 1+\frac{(\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6)(1-\theta)}{4(\theta\iota^3\sqrt{r}\kappa+(1-\theta))\|\mathbf{x}_1\|^6},\end{aligned}$$

where the last inequality follows from also

$$\|\mathbf{z}_k\|/\alpha_k^{(1)}=\left(\sum_{i=1}^r(\alpha_k^{(i)})^2\right)^{\frac{1}{2}}\Big/\alpha_k^{(1)}\leq\sqrt{r}\kappa$$

by (4.7). Therefore, for any $k\geq T_{1,1}$, we have

$$\begin{aligned}\rho_k&\geq\left(1+\frac{(\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6)(1-\theta)}{4(\theta\iota^3\sqrt{r}\kappa+(1-\theta))\|\mathbf{x}_1\|^6}\right)^{k-T_{1,1}}\rho_{T_{1,1}}\\&\overset{(4.9)}{\geq}\left(1+\frac{(\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6)(1-\theta)}{4(\theta\iota^3\sqrt{r}\kappa+(1-\theta))\|\mathbf{x}_1\|^6}\right)^{k-T_{1,1}}\frac{1}{\sqrt{r}\kappa}.\end{aligned}$$

Thus, for all k satisfying

$$k\geq T_{1,1}+\frac{\log\frac{2+\gamma}{\gamma}+\log(\sqrt{r}\kappa)}{\log\left(1+\frac{(\|\mathbf{x}_1\|^6-\|\mathbf{x}_2\|^6)(1-\theta)}{4(\theta\iota^3\sqrt{r}\kappa+(1-\theta))\|\mathbf{x}_1\|^6}\right)}:=T_{1,1}+T_{1,2},$$

we have $\rho_k\geq\frac{2+\gamma}{\gamma}$. Obviously,

$$T_{1,2}\leq C'_{\theta,\delta}\iota^3\sqrt{r}\kappa(\log\gamma^{-1}+\log r+\log\kappa)\leq C_{\theta,\delta}r^{2.1}\kappa^{4.2}(\log\gamma^{-1}+\log r+\log\kappa)$$

for some constants $C'_{\theta,\delta}$ and $C_{\theta,\delta}$.

By setting $T_1:=T_{1,1}+T_{1,2}$, we conclude the proofs of Step 3.

Step 4. Let T_2 be the minimum integer such that $T_2\geq T_1$ and

$$|\alpha^{(1)}_{T_2}-\|\mathbf{x}_1\||\leq\frac{\gamma}{2}\|\mathbf{x}_1\|.$$

We prove that

$$T_2\leq T_1+C_{\theta,\|\mathbf{x}_1\|}\gamma^{-1}(\log r+\log\kappa).$$

Noticing the definition

$$\nu_k:=\frac{\alpha^{(1)}_k}{\|\mathbf{x}_1\|},$$

it suffices to determine the minimum T_2 such that $T_2\geq T_1$ and $\nu_{T_2}\in\left[1-\frac{\gamma}{2},1+\frac{\gamma}{2}\right]$. By the result of Step 3, we have

$$\rho_k\geq\frac{2+\gamma}{\gamma}>\sqrt{\frac{2-\gamma}{\gamma}}$$

for all $k\geq T_1$, which implies

$$(1+\rho_k^{-2})^{-1}\geq 1-\frac{\gamma}{2}$$

for all $k\geq T_1$. This together with (4.11) gives

$$\nu_{k+1}\geq\left(\theta+(1-\theta)\left(1-\frac{\gamma}{2}\right)^2\nu_k^{-3}\right)\nu_k,\qquad\forall k\geq T_1. \tag{4.22}$$

We consider the following three cases.

- Case I: $\nu_{T_1}\in\left[1-\frac{\gamma}{2},1+\frac{\gamma}{2}\right]$. We are done.
- Case II: $\nu_{T_1}<1-\frac{\gamma}{2}$.

  It follows from (4.22) that $\nu_{k+1}>\nu_k$ if $\nu_k<1-\frac{\gamma}{2}$. In other words, the first few terms in $\nu_{T_1},\nu_{T_1+1},\cdots$ is monotonically strictly increasing. Let K_1 be the minimum integer such that $K_1>T_1$ and $\nu_{K_1}\geq 1-\frac{\gamma}{2}$. In the following we will show that $\nu_{K_1}\leq 1+\frac{\gamma}{2}$. Since

  $$\nu_{K_1}\leq\theta\nu_{K_1-1}+(1-\theta)\nu_{K_1-1}^{-2}:=f(\nu_{K_1-1}),$$

  We only need to show

  $$f(\nu_{K_1-1})\leq 1+\frac{\gamma}{2}.$$

We have $f'(t_0)=0$ if and only if

$$t_0=\left(\frac{2(1-\theta)}{\theta}\right)^{1/3},$$

and $f'(t)>0$ for $t\in[t_0,+\infty)$. Moreover, (4.22) yields

$$\begin{aligned}\nu_{K_1-1}\geq&\frac{\theta}{2}\nu_{K_1-2}+\frac{\theta}{2}\nu_{K_1-2}+(1-\theta)\left(1-\frac{\gamma}{2}\right)^2\nu_{K_1-2}^{-2}\\ \geq&1.88\theta^{\frac{2}{3}}(1-\theta)^{\frac{1}{3}}\left(1-\frac{\gamma}{2}\right)^{2/3}:=t_1.\end{aligned}$$

Let $t_2:=1-\frac{\gamma}{2}$. Therefore, provided $t_1\geq t_0$, i.e.,

$$\left(\frac{2(1-\theta)}{\theta}\right)^{1/3}\leq 1.88\theta^{\frac{2}{3}}(1-\theta)^{\frac{1}{3}}\left(1-\frac{\gamma}{2}\right)^{2/3}\iff\theta\geq\frac{2^{\frac{1}{3}}}{1.88}\left(1-\frac{\gamma}{2}\right)^{-2/3},\quad(4.23)$$

we have $t_2>\nu_{K_1-1}\geq t_1\geq t_0$ and thus

$$f(\nu_{K_1-1})\leq f(t_2).$$

A simple calculation reveals that

$$f(t_2)\leq 1+\frac{\gamma}{2}\iff\theta\geq\frac{\frac{1}{2}\gamma+\frac{1}{4}\gamma^2-\frac{1}{8}\gamma^3}{\frac{3}{2}\gamma-\frac{3}{4}\gamma^2+\frac{\gamma^3}{8}}.\quad(4.24)$$

Now we choose $\theta\in(0.7,1)$, so that both (4.23) and (4.24) are satisfied for all $\gamma\leq 0.05$. Consequently, $\nu_{K_1}\in\left[1-\frac{\gamma}{2},1+\frac{\gamma}{2}\right]$. It remains to estimate K_1. Using (4.22) and the lower bound of ν_k leads to,

$$\begin{aligned}1-\frac{\gamma}{2}>\nu_{K_1-1}\geq&\left(\theta+(1-\theta)\left(1-\frac{\gamma}{2}\right)^{-1}\right)\nu_{K_1-2}\geq\cdots\\ \geq&\left(\theta+(1-\theta)\left(1-\frac{\gamma}{2}\right)^{-1}\right)^{K_1-T_1}\nu_{T_1+1}\\ \geq&\left(\theta+(1-\theta)\left(1-\frac{\gamma}{2}\right)^{-1}\right)^{K_1-T_1}\cdot 1.88\theta^{\frac{2}{3}}(1-\theta)^{\frac{1}{3}}\left(1-\frac{\gamma}{2}\right)^{2/3},\end{aligned}$$

from which it follows that, for all $\gamma\in(0,0.05)$ and $\theta\in(0.7,1)$,

$$\begin{aligned}K_1-T_1\leq&\frac{\frac{1}{3}\log\left(1-\frac{\gamma}{2}\right)-\log\left(1.88\theta^{2/3}(1-\theta)^{1/3}\right)}{\log\left(\theta+(1-\theta)\left(1-\frac{\gamma}{2}\right)^{-1}\right)}\\ \leq&\frac{-\log\left(1.88\theta^{2/3}(1-\theta)^{1/3}\right)}{\frac{1}{2}(1-\theta)\frac{\gamma}{2-\gamma}}\leq C_\theta\gamma^{-1}\end{aligned}$$

for some constant C_θ. Therefore, we have $\nu_{K_1}\geq 1-\frac{\gamma}{2}$ with $K_1-T_1\leq C_\theta\gamma^{-1}$.

- Case III: $\nu_{T_1}>1+\frac{\gamma}{2}$. It follows from (4.13) that

$$\nu_{k+1}\leq\left(\theta+(1-\theta)\nu_k^{-3}\right)\nu_k<\nu_k \qquad \text{for } k=T_1,T_1+1,\cdots,K_2-1, \tag{4.25}$$

where we have defined K_2 to be the minimum integer such that $K_2>T_1$ and $\nu_{K_2}\leq 1+\frac{\gamma}{2}$. If $\nu_{K_2}\geq 1-\frac{\gamma}{2}$, then $\nu_{K_2}\in\left[1-\frac{\gamma}{2},1+\frac{\gamma}{2}\right]$ and we are done. Otherwise, $\nu_{K_2}<1-\frac{\gamma}{2}$, and it is reduced to Case II; more precisely, $\nu_{K_3}\in\left[1-\frac{\gamma}{2},1+\frac{\gamma}{2}\right]$ with $K_3-K_2\leq C_\theta\gamma^{-1}$.

It remains to estimate K_2. Eq. (4.25) implies

$$\begin{aligned}
1+\frac{\gamma}{2}<\nu_{K_2-1}&\leq\left(\theta+(1-\theta)\nu_{K_2-2}^{-3}\right)\nu_{K_2-2}\leq\left(\theta+(1-\theta)\left(1+\frac{\gamma}{2}\right)^{-3}\right)\nu_{K_2-2}\\
&\leq\cdots\leq\left(\theta+(1-\theta)\left(1+\frac{\gamma}{2}\right)^{-3}\right)^{K_2-T_1-1}\cdot\nu_{T_1}\\
&\leq\left(\theta+(1-\theta)\left(1+\frac{\gamma}{2}\right)^{-3}\right)^{K_2-T_1-1}\cdot\nu
\end{aligned}$$

with ν being the upper bound of ν_k defined in (4.10). Therefore, for any $\gamma\in(0,0.05)$,

$$\begin{aligned}
K_2-T_1&\leq 1+\frac{\log\frac{\nu}{1+\frac{\gamma}{2}}}{-\log\left(\theta+(1-\theta)\left(1+\frac{\gamma}{2}\right)^{-3}\right)}\leq 1+\frac{\log\nu}{(1-\theta)\left(1-\left(1+\frac{\gamma}{2}\right)^{-3}\right)}\\
&=1+\frac{\log\nu}{(1-\theta)\left(1-\left(1+\frac{\gamma}{2}\right)^{-1}\right)\left(1+\left(1+\frac{\gamma}{2}\right)^{-1}+\left(1+\frac{\gamma}{2}\right)^{-2}\right)}\\
&\leq C'_{\theta,\|\mathbf{x}_1\|}\gamma^{-1}(\log r+\log\kappa).
\end{aligned}$$

Combining all the cases above, we have

$$\begin{aligned}
T_2&\leq\max\{K_1,K_2,K_3\}\\
&\leq T_1+C_\theta\gamma^{-1}+C'_{\theta,\|\mathbf{x}_1\|}\gamma^{-1}(\log r+\log\kappa)\\
&\leq T_1+C_{\theta,\|\mathbf{x}_1\|}\gamma^{-1}(\log r+\log\kappa).
\end{aligned}$$

Step 5. We prove

$$\sqrt{\sum_{i=2}^{r}\left(\alpha_i^{(i)}\right)^2+\beta_k^2}\leq\frac{\gamma}{2}\|\mathbf{x}_1\|$$

for $k=T_2$.

Since

$$\rho_k \geq \frac{2+\gamma}{\gamma} \quad \text{and} \quad \alpha_k^{(1)} \leq \left(1+\frac{\gamma}{2}\right)\|\mathbf{x}_1\|,$$

when $k=T_2$, we have

$$\frac{(1+\frac{\gamma}{2})\|\mathbf{x}_1\|}{\sqrt{\sum_{i=2}^r \left(\alpha_k^{(i)}\right)^2+\beta_k^2}} \geq \frac{\alpha_k^{(1)}}{\sqrt{\sum_{i=2}^r \left(\alpha_k^{(i)}\right)^2+\beta_k^2}} = \rho_k \geq \frac{2+\gamma}{\gamma},$$

which implies

$$\sqrt{\sum_{i=2}^r \left(\alpha_i^{(i)}\right)^2+\beta_k^2} \leq \frac{\gamma}{2}\|\mathbf{x}_1\|.$$

To sum up: when $k=T_2$, we have

$$\begin{aligned}\|\mathbf{z}_k-\mathbf{x}_1\| &= \sqrt{\left|\alpha_k^{(1)}-\|\mathbf{x}_1\|\right|^2+\sum_{i=2}^r \left(\alpha_i^{(i)}\right)^2+\beta_k^2} \\ &\leq \left|\alpha_k^{(1)}-\|\mathbf{x}_1\|\right|+\sqrt{\sum_{i=2}^r \left(\alpha_i^{(i)}\right)^2+\beta_k^2} \leq \gamma\|\mathbf{x}_1\|.\end{aligned}$$

Moreover, T_2 satisfies

$$T_2 \leq C_{\theta,\delta} r^{2.1}\kappa^{4.2}(\log\gamma^{-1}+\log r+\log\kappa)+C_{\theta,\|\mathbf{x}_1\|}\gamma^{-1}(\log r+\log\kappa).$$

Thus, we completes the proof. □

5 Numerical implements

In this section, we explore the performance of our proposed method for tensor decomposition. For the stepsize, we initially set it to $\mu_0=10^{-14}$ and then use Barzilai-Borwein's method (B-B method) to choose it during the iteration. Here we choose stepsize as the absolute value of itself because stepsize is positive in general. The algorithm stops when $\|\mathbf{g}_k\|<10^{-16}$, where $\mathbf{g}_k=\nabla f(\mathbf{z}_k)-\nabla f(\mathbf{z}_{k+1})$, or the maximal iteration number is reached.

5.1 Numerical results

We demonstrate the numerical performance of our proposed method to find the best rank-1 approximation. In order to verify the global convergence, we generate

$$\mathcal{A}=\sum_{i=1}^{r}\mathbf{x}_i\otimes\mathbf{x}_i\otimes\mathbf{x}_i,$$

where $\{\mathbf{x}_i\}_{i=1}^r$ are nonzero orthogonal vectors and $\|\mathbf{x}_1\|>\|\mathbf{x}_2\|\geq\cdots\geq\|\mathbf{x}_r\|>0$. We consider the following two examples.

Example 5.1. Let

$$\mu=(\mu_1\ \cdots\ \mu_6)=(0.6859\ -0.5641\ -0.5491\ 0.3792\ -0.3530\ 0.0945),$$

$$\mathbf{y}=[\mathbf{y}_1,\cdots,\mathbf{y}_6]=\begin{bmatrix}-0.3628 & -0.2999 & 0.2888 & -0.2649 & -0.1895 & -0.2424\\ 0.3063 & -0.4028 & 0.1969 & -0.3806 & 0.5056 & 0.1386\\ -0.5211 & -0.3685 & -0.1368 & 0.1244 & -0.1636 & -0.3962\\ 0.3728 & -0.2374 & -0.2764 & 0.0824 & -0.6928 & 0.2981\\ 0.0635 & -0.3267 & -0.6741 & -0.4802 & 0.0819 & -0.1060\\ 0.3692 & -0.3637 & 0.5396 & -0.0581 & -0.2898 & -0.2125\\ 0.2970 & -0.3365 & -0.1774 & 0.6795 & 0.3132 & -0.3435\\ -0.3681 & -0.4511 & 0.0750 & 0.2589 & 0.1135 & 0.7084\end{bmatrix}.$$

Set $\mathbf{x}_i=\sqrt[3]{\mu_i}\mathbf{y}_i$, $i=1,\cdots,6$, and

$$\mathcal{A}=\sum_{i=1}^{6}\mathbf{x}_i^{\otimes 3}.$$

Example 5.2. Let

$$\mathcal{A}=\mathbf{x}_1\otimes\mathbf{x}_1\otimes\mathbf{x}_1+10\cdot\mathbf{x}_2\otimes\mathbf{x}_2\otimes\mathbf{x}_2,$$

where $\mathbf{x}_1,\mathbf{x}_2\in\mathbb{R}^n$, $\mathbf{x}_1\sim\mathcal{N}(\mathbf{0},I)$ and $\mathbf{x}_2^T\mathbf{x}_1=0$, $\|\mathbf{x}_2\|=1$.

We first evaluate the effectiveness of our proposed algorithm in terms of the smallest L required for successful tensor decomposition. We use 100 trials for test. In each trial, we generate L random sampling vectors for initialization. We declare it is successful if the error

$$\left\|\mathbf{z}-\frac{\mathbf{z}^T\mathbf{x}_1}{\|\mathbf{x}_1\|^2}\mathbf{x}_1\right\|<10^{-5},$$

where $\mathbf{z}$ is the numerical solution. The empirical probability of success is defined as the average of success over 100 trials. We use $n=128$ for Example 5.2. Fig. 1 gives

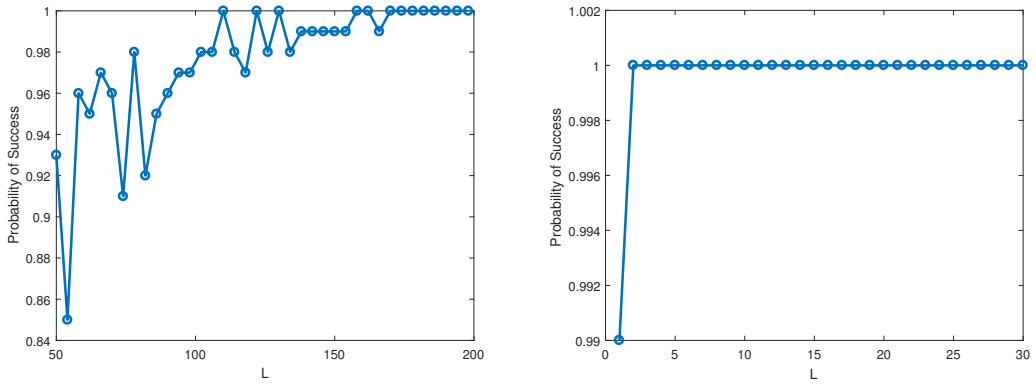

Figure 1: The plot of the empirical probability of success against the over sampling number L, left for Example 5.1 and right for Example 5.2.

the plot of the empirical probability of success against the over sampling number L. We see that for a 100% successful convergence to global minimizer we need a large L–L is about 160 for Example 5.1. We see also that the success rate increases with respect to L. It confirms that a multi-start of our algorithm with large L is necessary for a global minimum.

Next, we demonstrate the efficiency of our proposed method. To investigate the convergence of $\mathbf{z}_k$ generated by the proposed algorithm, we calculate and demonstrate the change of the relative error $\|\mathbf{z}_k-\mathbf{x}_1\|/\|\mathbf{x}_1\|$ to the global minimizer, the coefficient

$$\alpha_k^{(1)}=\frac{\langle \mathbf{z}_k,\mathbf{x}_1\rangle}{\|\mathbf{x}_1\|}$$

of the projection of $\mathbf{z}_k$ onto the span of $\mathbf{x}_1$, and the coefficient

$$\chi_k=\left(\|\mathbf{z}_k\|^2-(\alpha_k^{(1)})^2\right)^{1/2}$$

of $\mathbf{z}_k$ onto the orthogonal complementary of $\mathbf{x}_1$. We use $L=200$ in the following all experiments. For Example 5.2, we actually vary the dimension numbers n with $n=64,128,256,512$ and the results are almost the same. Here we only show the results with $n=128$ to save page.

In Fig. 2, we plot $\|\mathbf{z}_k-\mathbf{x}_1\|/\|\mathbf{x}_1\|$, $|\alpha_k^{(1)}|$, and χ_k versus the iteration number k for Examples 5.1 and 5.2, respectively.

From these figures, we observe that the gradient descent algorithm with the initialization in (3.2) to solve (1.4) does not hit the saddle points, local minima, or $\mathbf{0}$ at all. It will converge to global minimizer, although the nonconvex optimization objective function here has numerous critical points, including local (global) minima,

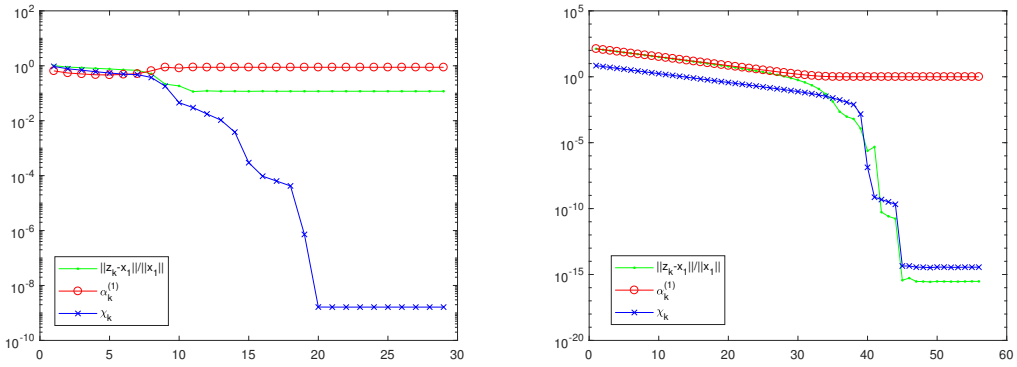

Figure 2: Plots of $\|\mathbf{z}_k-\mathbf{x}_1\|/\|\mathbf{x}_1\|$, $|\alpha_k^{(1)}|$, and χ_k versus the iteration number, left for Example 5.1 and right for Example 5.2.

saddle points, and $\mathbf{0}$. That is, starting from the designed initialization, gradient descent will bypass the saddle points and entry the local region containing the truth. This observation is consistent with our theoretical results in the previous sections.

6 Conclusions

In this paper, we use nonconvex optimization model (1.4) to find the best rank-one approximation of a symmetric tensor. First of all, we give the optimization landscape of the nonconvex optimization (1.4). We find that the local minimizers are the factors in the CP low-rank decomposition of the given symmetric tensor. We use gradient descent to solve the nonconvex optimization model. We prove that gradient descent from one careful initialization will arrive at one convex region of the global minimizer after at most finite steps and then converge to the local minimizer linearly. Numerical results coincide with the theoretical proof. For the future directions of tensor decomposition, we may focus on the theoretical proof of noisy tensor and some applications of tensor decomposition, for example, ℓ^p-norm maximization.

References

[1] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, Tensor decompositions for learning latent variable models, J. Machine Learn Res., 15 (2014), pp. 2773–2832.

[2] K. Batselier and N. Wong, Symmetric tensor decomposition by an iterative eigendecomposition algorithm, J. Comput. Appl. Math., (2016).
[3] E. J. Candès, X. Li, and M. Soltanolkotabi, Phase retrieval via wirtinger flow: Theory and algorithms, IEEE Trans. Info. Theory, 61 (2015), pp. 1985–2007.
[4] J.-F. Cardoso, Blind signal separation: statistical principles, Proceedings of the IEEE, 86 (1998), pp. 2009–2025.
[5] Y. Chen, Y. Chi, J. Fan, and C. Ma, Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval, Math. Program., (2018), pp. 1–33.
[6] A. Cichocki and S.-I. Amari, Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications, vol. 1, John Wiley & Sons, 2002.
[7] P. Comon and M. Rajih, Blind identification of under-determined mixtures based on the characteristic function, Signal Process., 86 (2006), pp. 2271–2281.
[8] L. De Lathauwer, B. De Moor, and J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.
[9] D. L. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, IEEE Trans. Info. Theory, 47 (2001), pp. 2845–2862.
[10] A. T. Erdogan, On the convergence of ICA algorithms with symmetric orthogonalization, IEEE Trans. Signal Process., 57 (2009), pp. 2209–2221.
[11] B. Jiang, S. Ma, and S. Zhang, Tensor principal component analysis via convex optimization, Math. Program., 150 (2015), pp. 423–457.
[12] E. Kofidis and P. A. Regalia, On the best rank-1 approximation of higher-order supersymmetric tensors, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 863–884.
[13] T. G. Kolda, Symmetric orthogonal tensor decomposition is trivial, arXiv preprint arXiv:1503.01375, (2015).
[14] T. G. Kolda and B. W. Bader, Tensor decompositions and applications, SIAM Rev., 51 (2009), pp. 455–500.
[15] T. G. Kolda and J. R. Mayo, Shifted power method for computing tensor eigenpairs, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 1095–1124.
[16] J. B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, Linear Algebra Appl., 18 (1977), pp. 95–138.
[17] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, Gradient descent only converges to minimizers, Conference on Learning Theory, (2016), pp. 1246–1257.
[18] L.-H. Lim, Singular values and eigenvalues of tensors: a variational approach, 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, IEEE, (2005), pp. 129–132.
[19] H. Liu, Symmetric tensor decomposition by alternating gradient descent, Numer. Linear Algebra Appl., online, (2021).
[20] J. Pan and M. K. Ng, Symmetric orthogonal approximation to symmetric tensors with applications to image reconstruction, Numer. Linear Algebra Appl., 25 (2018), p. e2180.

[21] L. Qi, Eigenvalues of a real supersymmetric tensor, J. Symbolic Comput., 40 (2005), pp. 1302–1324.
[22] L. Qi, F. Wang, and Y. Wang, Z-eigenvalue methods for a global polynomial optimization problem, Math. Program., 118 (2009), pp. 301–316.
[23] P. A. Regalia and E. Kofidis, Monotonic convergence of fixed-point algorithms for ICA, IEEE Trans. Neural Networks, 14 (2003), pp. 943–949.
[24] Y. Shen, Y. Xue, J. Zhang, K. Letaief, and V. Lau, Complete dictionary learning via ℓ^p-norm maximization, Conference on Uncertainty in Artificial Intelligence, PMLR, (2020), pp. 280–289.
[25] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, Parallel factor analysis in sensor array processing, IEEE Trans. Signal Process., 48 (2000), pp. 2377–2388.
[26] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, arXiv preprint arXiv:1011.3027, (2010).
[27] Y. Zhai, Z. Yang, Z. Liao, J. Wright, and Y. Ma, Complete dictionary learning via ℓ^4-norm maximization over the orthogonal group., J. Mach. Learn. Res., 21 (2020), pp. 1–68.