CONVERGENCE RATE ANALYSIS FOR THE HIGHER ORDER POWER METHOD IN BEST RANK ONE APPROXIMATIONS OF TENSORS

SHENGLONG HU AND GUOYIN LI

Abstract. A popular and classical method for finding the best rank one approximation of a real tensor is the higher order power method (HOPM). It is known in the literature that the iterative sequence generated by HOPM converges globally, while the convergence rate can be superlinear, linear or sublinear. In this paper, we examine the convergence rate of HOPM in solving the best rank one approximation problem of real tensors. We first show that the iterative sequence of HOPM always converges globally and provide an explicit eventual sublinear convergence rate. The sublinear convergence rate estimate is in terms of the dimension and the order of the underlying tensor space. Then, we examine the concept of nondegenerate singular vector tuples and show that, if the sequence of HOPM converges to a nondegenerate singular vector tuple, then the global convergence rate is R-linear. We show that, for almost all tensors (in the sense of Lebesgue measure), all the singular vector tuples are nondegenerate, and so, the HOPM "typically" exhibits global R-linear convergence rate. Moreover, without any regularity assumption, we establish that the sequence generated by HOPM always converges globally and R-linearly for orthogonally decomposable tensors with order at least 3. We achieved this by showing that each nonzero singular vector tuple of an orthogonally decomposable tensor with order at least 3 is nondegenerate.

1. Introduction

As generalizations of matrices, tensors (a.k.a. hypermatrices or multi-way arrays) are ubiquitous and inevitable in modeling and scientific computing in applied sciences [28, 30, 36, 37]. Among the many aspects on the developments of tensors in recent years, tensor approximations/decompositions and their related topics have been becoming the focus, see [8, 14, 35, 36] and references therein. The applications of these techniques are diverse and broad, including computational complexity [37], data analysis [11], pattern recognition [2], principal component analysis [9, 12, 13], scientific computing [1, 10], signal processing [16], etc. We refer interested readers to the surveys [29, 35] and books [30, 37], and references therein for more details.

One of the important problems in tensor approximation and tensor decomposition areas is the best rank one approximation problem which has many important applications in signal processing [9, 19]. Mathematically, the best rank one approximation of a tensor is to find a rank one tensor so that the deviation (or approximation error) computed by subtracting the rank one approximation from the given tensor has minimum Hilbert-Schmidt norm (minimization formulation, cf. (5)). It has an equivalent maximization problem formulation [18] (cf. (6)). Unlike many tensor approximation or decomposition problems which are in general ill-posed and computationally hard problems [20,31], the best rank one approximation problems of tensors are always well-posed [18,20]. Due to its importance, the best rank one approximations problem of tensors has attracted a lot of attentions. The literature of this problem is vast, and ranges from algorithm developments and their convergence properties [18, 33, 58–60, 63] to applications [1, 18, 20]. An important class of structured tensors which deserves special attentions is the so-called *orthogonally decomposable tensors*. This class of tensors arises from machine learning with latent variables [2], blind source separation [12] and statistics [13], and admits many highly desired features such as *unique rank*

2010 *Mathematics Subject Classification.* 15A18; 15A69; 65F18.

Key words and phrases. Tensor, best rank one approximation, Łojasiewicz's inequality, sublinear convergence, singular vector, nondegenerate, R-linear convergence.

decompositions which distinguishes from general tensors. Therefore, understanding orthogonally decomposable tensors has become an important research direction from both theoretical and computational point of view [34, 63]. As we will see later, the theory developed in this article will also offer new insights to this important class of tensors.

A popular and classical method for finding the best rank one approximation of a real tensor is the higher order power method (HOPM). To the best knowledge of the authors, the higher order power method (HOPM) for solving the best rank one approximation problem based on the maximization formulation was firstly proposed by De Lathauwer, De Moor, and Vandewalle in [18]. The convergence properties for this method in the symmetric tensor cases were studied extensively by Kofidis and Regalia [33] under a convexity assumption. The higher order power method in [18] is an application of the classical nonlinear block Gauss-Seidel (coordinate descent) method [49] to the maximization formulation of the best rank one approximation problem of tensors. When the block Gauss-Seidel (coordinate descent) method is applied to the equivalent minimization formulation, it yields the so-called alternating least square method (ALS). As pointed out in [59], these two methods are actually equivalent in the sense that they will generate the same iterative sequence up to scaling given the same initialization. Mohlenkamp [45] studied a general alternating least square method and show that any limiting point of the iterative sequence generated by this method is a singular vector tuple of the tensor [43]. Later on, Wang and Chu [60] gave a rigorous analysis establishing the global convergence of the iterative sequence for generic tensors. The global convergence is completely solved by Uschmajew [59], who proved that the global convergence occurs without any assumption by using Łojasiewicz's inequality. He also pointed out that a local convergence rate could be given if one knows the Łojasiewicz exponent of the objective function which is however hard to estimate or unknown in general [3, 5].

The convergence rate for the HOPM in solving best rank one approximation problem of tensors is much more subtle. In the literature, Zhang and Golub [63] studied the local quadratic convergence of a Newton method under a nonsingularity assumption, while there is no theoretical conclusion on which class of tensors satisfying the nonsingularity assumption. Uschmajew [58] also established a locally linear convergence result under an assumption on the rank of the Hessian matrix of the objective function based on a variation of the minimization formulation for HOPM/ALS. Interestingly, this assumption is fulfilled if the Euclidean distance between the given tensor $\mathcal{A}$ (with order d) and the set of rank one tensors is strictly smaller than $\frac{\|\mathcal{A}\|_{\mathrm{F}}}{\sqrt{3-2/d}}$. This is equivalent to saying that the best approximation ratio of this tensor in the sense of [51] is no smaller than $\sqrt{\frac{2d-2}{3d-2}}$. Unfortunately, in general, the best approximation ratio also depends on the dimensions of the tensor space and is typically much smaller than $\sqrt{\frac{2d-2}{3d-2}}$ [51]. Very recently, Espig and Khachatryan [24], and Espig, Hackbusch and Khachatryan [23] thoroughly examined the behavior of HOPM/ALS and gave a detailed analysis analyzing the global convergence rate, under a suitable regularity assumption. In particular, they provided concrete examples as well as classes of structured tensors, and showed that the global convergence rate of HOPM/ALS can be superlinear, linear or sublinear under the prescribed regularity assumption. Thus, one cannot always expect an overall linearly convergent phenomena.

Despite the above significant contributions, there are still several important questions need to be answered. For example, can we establish the sublinear convergence with an explicit overall global convergence rate estimate for HOPM in solving best rank one approximation problems for tensors? Can we derive the linear convergence of HOPM for almost all tensors[1]? Can we identify certain important classes of tensors so that, without assuming any regularity assumption, HOPM always exhibits linear convergence for this particular classes of tensors?

[1] Here, a property holds for almost all tensors means that the set of tensors for which the property does not hold is a set of Lebesgue measure zero.

The purpose of this paper is to examine the convergence rate for HOPM in solving the best rank one approximation for real tensors, and provide answers for the above questions. In particular, the main contributions of this article are

(1) We establish an overall sublinear convergence of HOPM in solving the best rank one approximation for real tensors, and provide *an explicit sublinear convergence rate* in terms of the dimension of the underlying space and the order of the tensor (Theorem 3.2). Interestingly, the derived convergence rate estimate of the objective value sequence is sharper than the usual convergence rate established for first-order algorithms in optimization literature.

(2) Then, we examine the concept of nondegenerate singular vector tuples and show that, if the sequence of HOPM converges to a nondegenerate singular vector tuple, then the global convergence rate is R-linear. Our result does not require the limit point to be a local/global solution of the corresponding optimization formulation, and so, distinguishes with the sufficient conditions for linear convergence in the literature (e.g. [58, 63]). More importantly, we further justify that, for almost all tensors, each of its singular vector tuples is nondegenerate. As a result, the HOPM *typically exhibits global R-linear convergence rate* (Theorem 4.4 and Corollary 5.4).

(3) We further demonstrate that, without any regularity assumptions, HOPM always converges R-linearly for orthogonally decomposable tensors with order at least 3 by establishing the fact that every nonzero singular vector tuple of an orthogonally decomposable tensor with order at least 3 is nondegenerate (Proposition 6.5 and Corollary 6.6).

The organization of the rest paper will be as follows. In Section 2, we present preliminaries on the notation and specific tools that will be used in the later sections including basic operations on tensors, singular vectors/values of tensors, as well as Łojasiewicz's inequalities. In Section 3, we establish the overall sublinear convergence rate of HOPM for best rank one approximations to tensors. In Section 4, we obtain that, if the sequence of HOPM converges to a nondegenerate singular vector tuple, then the global convergence rate is R-linear. In Section 5, we show that HOPM exhibits R-linear convergence for almost all tensors. We achieve this by showing that, for almost all tensors, each of its singular vector tuple is nondegenerate. In Section 6, we further study nondegenerate singular vector tuples, showing that for orthogonally decomposable tensors, each nonzero singular vector tuple is nondegenerate, and thus the HOPM always exhibits an R-linear convergence rate.

2. Preliminaries

In this section, preliminaries will be given for the subsequent analysis. Unless otherwise stated, the focus will be real tensors in the given tensor space $\mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$ for some positive integers $d \geq 3$ and $n_1, \dots, n_d \geq 2$.

2.1. Notation and mappings.

Throughout this paper, $\|\cdot\|$ is reserved for the Euclidean norm of a vector. Given a *block vector*

$$\mathbf{x} := (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d} \simeq \mathbb{R}^{n_1 + \cdots + n_d} \text{ with } \mathbf{x}_i \in \mathbb{R}^{n_i} \text{ for all } i = 1, \dots, d,$$

we define a mapping $\tau : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d} \to \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$ as the rank one tensor with order d defined by $\{\mathbf{x}_1, \dots, \mathbf{x}_d\}$, that is,

$$\tau(\mathbf{x}) = \mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_d. \tag{1}$$

This mapping is well-known as Segre mapping or Segre embedding [56]. For each $i \in \{1, \dots, d\}$, the mapping $\tau_i : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d} \to \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_{i-1}} \otimes \mathbb{R}^{n_{i+1}} \otimes \cdots \otimes \mathbb{R}^{n_d}$ is defined as

$$\tau_i(\mathbf{x}) = \mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{i-1} \otimes \mathbf{x}_{i+1} \otimes \cdots \otimes \mathbf{x}_d.$$

It is the rank one tensor defined by $\{\mathbf{x}_1,\dots,\mathbf{x}_d\}$ with $\mathbf{x}_i$ removed. Given two tensors $\mathcal{A},\mathcal{B}\in\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$ with order d and the entries being indexed as $a_{i_1\dots i_d}$ and $b_{i_1\dots i_d}$ respectively, the inner product is defined as

$$\langle\mathcal{A},\mathcal{B}\rangle := \sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d} a_{i_1\dots i_d}b_{i_1\dots i_d},$$

with the corresponding induced norm given by $\|\mathcal{A}\|_{\mathrm{HS}} := \sqrt{\langle\mathcal{A},\mathcal{A}\rangle}$. This norm is a generalization of the matrix Frobenius norm and is termed as the Hilbert-Schmidt norm. Direct calculation gives us that

$$\langle\mathcal{A},\tau(\mathbf{x})\rangle = \sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d} a_{i_1\dots i_d}(\mathbf{x}_1)_{i_1}\dots(\mathbf{x}_d)_{i_d}.$$

For the sake of notational convention, we also write $\mathcal{A}\tau_i(\mathbf{x})$ as a vector in $\mathbb{R}^{n_i}$ with its j-th component being

$$\sum_{k_1=1}^{n_1}\cdots\sum_{k_{i-1}=1}^{n_{i-1}}\sum_{k_{i+1}=1}^{n_{i+1}}\cdots\sum_{k_d=1}^{n_d} a_{k_1\dots k_{i-1}jk_{i+1}\dots k_d}(\mathbf{x}_1)_{k_1}\dots(\mathbf{x}_{i-1})_{k_{i-1}}(\mathbf{x}_{i+1})_{k_{i+1}}\dots(\mathbf{x}_d)_{k_d}.$$

Direct verification shows that, for each $i=1,\dots,d$,

$$\big(\mathcal{A}\tau_i(\mathbf{x})\big)^{\mathsf{T}}\mathbf{x}_i = \langle\mathcal{A},\tau(\mathbf{x})\rangle. \tag{2}$$

If $i<j$, we define $\mathcal{A}\tau_{i,j}(\mathbf{x})$ as a matrix in $\mathbb{R}^{n_i\times n_j}$ whose (s,t)-th component is given as

$$\begin{aligned}&\langle\mathcal{A},\mathbf{x}_1\otimes\cdots\otimes\mathbf{e}^i_s\otimes\cdots\otimes\mathbf{e}^j_t\otimes\cdots\otimes\mathbf{x}_d\rangle\\ =&\sum_{k_1=1}^{n_1}\cdots\sum_{k_{i-1}=1}^{n_{i-1}}\sum_{k_{i+1}=1}^{n_{i+1}}\cdots\sum_{k_{j-1}=1}^{n_{j-1}}\sum_{k_{j+1}=1}^{n_{j+1}}\cdots\sum_{k_d=1}^{n_d}\\ &a_{k_1\dots k_{i-1}sk_{i+1}\dots k_{j-1}tk_{j+1}\dots k_d}(\mathbf{x}_1)_{k_1}\dots(\mathbf{x}_{i-1})_{k_{i-1}}(\mathbf{x}_{i+1})_{k_{i+1}}\dots(\mathbf{x}_{j-1})_{k_{j-1}}(\mathbf{x}_{j+1})_{k_{j+1}}\dots(\mathbf{x}_d)_{k_d},\end{aligned}$$

where $\mathbf{e}^i_s\in\mathbb{R}^{n_i}$ is the s-th coordinate vector for all $s\in\{1,\dots,n_i\}$ and $i\in\{1,\dots,d\}$, and $\mathcal{A}\tau_{i,j}(\mathbf{x})$ is the transpose of the matrix $\mathcal{A}\tau_{j,i}(\mathbf{x})$ if $i>j$.

Let $\mathbb{S}^{n-1} := \{\mathbf{z}\in\mathbb{R}^n \mid \|\mathbf{z}\|=1\}$ be the unit sphere in $\mathbb{R}^n$. The following simple lemma shows that the function τ is Lipschitz continuous on the joint sphere $\mathbb{S} := \mathbb{S}^{n_1-1}\times\cdots\times\mathbb{S}^{n_d-1}$.

Lemma 2.1. *For any* $\mathbf{x},\mathbf{y}\in\mathbb{S}^{n_1-1}\times\cdots\times\mathbb{S}^{n_d-1}$*, we have*

$$\|\tau(\mathbf{x})-\tau(\mathbf{y})\|_{\mathrm{HS}} \le \sqrt{d}\|\mathbf{x}-\mathbf{y}\|.$$

Proof. Let $\mathbf{x},\mathbf{y}\in\mathbb{S}^{n_1-1}\times\cdots\times\mathbb{S}^{n_d-1}$. Note that $\sum_{i=1}^d\|\mathbf{x}_i-\mathbf{y}_i\|\le\sqrt{d}\|\mathbf{x}-\mathbf{y}\|$. To see the conclusion, it suffices to show that

$$\|\tau(\mathbf{x})-\tau(\mathbf{y})\|_{\mathrm{HS}} \le \sum_{i=1}^d\|\mathbf{x}_i-\mathbf{y}_i\|. \tag{3}$$

We now establish (3) by the method of mathematical induction on d. Clearly, the conclusion is trivially true in the case $d=1$. Now, suppose the conclusion is true for $d=l-1$. We now consider the case for $d = l$. Then, it follows from triangle inequality that

$$\begin{aligned}\|\tau(\mathbf{x}) - \tau(\mathbf{y})\|_{\mathrm{HS}} &= \|(\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{l-1} \otimes \mathbf{x}_l) - (\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{l-1} \otimes \mathbf{y}_l) \\ &\quad + (\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{l-1} \otimes \mathbf{y}_l) - (\mathbf{y}_1 \otimes \cdots \otimes \mathbf{y}_{l-1} \otimes \mathbf{y}_l)\|_{\mathrm{HS}} \\ &\le \|(\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{l-1} \otimes \mathbf{x}_l) - (\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{l-1} \otimes \mathbf{y}_l)\|_{\mathrm{HS}} \\ &\quad + \|(\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{l-1} \otimes \mathbf{y}_l) - (\mathbf{y}_1 \otimes \cdots \otimes \mathbf{y}_{l-1} \otimes \mathbf{y}_l)\|_{\mathrm{HS}} \\ &= \|\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{l-1}\|_{\mathrm{HS}} \|\mathbf{x}_l - \mathbf{y}_l\| + \|(\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{l-1}) - (\mathbf{y}_1 \otimes \cdots \otimes \mathbf{y}_{l-1})\|_{\mathrm{HS}} \|\mathbf{y}_l\| \\ &= \|\mathbf{x}_l - \mathbf{y}_l\| + \|\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_{l-1} - \mathbf{y}_1 \otimes \cdots \otimes \mathbf{y}_{l-1}\|_{\mathrm{HS}} \\ &\le \sum_{i=1}^{l} \|\mathbf{x}_i - \mathbf{y}_i\|,\end{aligned}$$

where the third equality follows from the spherical constraint, and the last inequality follows from the induction hypothesis. So, (3) holds, and hence, the conclusion follows. □

Similarly, the function τ_i is also Lipschitz continuous on the joint sphere.

2.2. **Singular vectors.** We first give the definitions of singular vector tuples.

Definition 2.2. *Given a tensor $\mathcal{A} \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$, a vector tuple $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathbb{S} = \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$ is called a (real) singular vector tuple of $\mathcal{A}$ if it is a critical point of the smooth function $G(\mathbf{x}) := \langle \mathcal{A}, \tau(\mathbf{x}) \rangle$ on the joint sphere $\mathbb{S}$. The value of G at a singular vector tuple is called a singular value. The corresponding vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_d\}$ are called singular vectors.*

The definitions of singular values/vectors were proposed by Lim [43] (a similar treatment for eigenvalues/eigenvectors was independently introduced by Qi [50]). Given a tensor $\mathcal{A} \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$, we note that, by using Lagrange multiplier method, the definition entails that a *singular vector tuple* $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_d) \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$ of $\mathcal{A}$ and a corresponding *singular value* σ satisfy

$$\begin{cases} \mathcal{A}\tau_1(\mathbf{x}) = \sigma \mathbf{x}_1, \\ \qquad \vdots \\ \mathcal{A}\tau_d(\mathbf{x}) = \sigma \mathbf{x}_d. \end{cases} \tag{4}$$

It is easy to see from (4) that

$$\sigma = \langle \mathcal{A}, \tau(\mathbf{x}) \rangle$$

for a singular vector tuple $\mathbf{x}$ and the corresponding singular value σ. Direct verification shows that whenever $(\mathbf{x}_1, \dots, \mathbf{x}_d)$ is a singular vector tuple with singular value σ, then $(\epsilon_1 \mathbf{x}_1, \dots, \epsilon_d \mathbf{x}_d)$ is also a singular vector tuple of the same singular value for any choices of $\epsilon_i \in \{-1, 1\}$ as long as $\prod_{i=1}^{d} \epsilon_i = 1$. This is usually viewed as one (class) when we counting the number of singular vector tuples. Another important symmetric property of the singular vector tuples is that $(\mathbf{x}_1, \dots, \mathbf{x}_d)$ is a singular vector tuple with singular value σ, if and only if $(\epsilon_1 \mathbf{x}_1, \dots, \epsilon_d \mathbf{x}_d)$ is a singular vector tuple with singular value $-\sigma$ for any choices of $\epsilon_i \in \{-1, 1\}$ with $\prod_{i=1}^{d} \epsilon_i = -1$. Thus, for many situations, one can simply focus on the singular vector tuples with nonnegative singular values (as in the classical matrix cases).

The next proposition follows from the compactness of the joint sphere and standard optimality condition theory [54], please see also [63] for the essential proof when $d = 3$.

Proposition 2.3. *Let $\mathcal{A} \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$. Then $\mathcal{A}$ has at least one singular vector tuple.*

2.3. **Best rank one approximation.** Given a tensor $\mathcal{A}\in\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$, the well-known best rank one approximation problem can be formulated as (cf. [18,63])

$$\min\left\{\|\mathcal{A}-\lambda\mathbf{x}_1\otimes\cdots\otimes\mathbf{x}_d\|_{\mathrm{HS}}^2 \mid \lambda\in\mathbb{R},\ \mathbf{x}=(\mathbf{x}_1,\dots,\mathbf{x}_d)\in\mathbb{S}^{n_1-1}\times\cdots\times\mathbb{S}^{n_d-1}\right\}. \tag{5}$$

Direct calculation shows that the best λ for the above problem is $\langle\mathcal{A},\tau(\mathbf{x})\rangle$ for any given $\mathbf{x}$. Thus, at a minimizer, we have

$$\|\mathcal{A}-\lambda\mathbf{x}_1\otimes\cdots\otimes\mathbf{x}_d\|_{\mathrm{HS}}^2=\|\mathcal{A}\|_{\mathrm{HS}}^2-\langle\mathcal{A},\tau(\mathbf{x})\rangle^2.$$

Hence, (5) is equivalent to

$$\max\left\{\langle\mathcal{A},\tau(\mathbf{x})\rangle \mid \mathbf{x}=(\mathbf{x}_1,\dots,\mathbf{x}_d)\in\mathbb{S}^{n_1-1}\times\cdots\times\mathbb{S}^{n_d-1}\right\}. \tag{6}$$

The next proposition can be proved with the spirit of Proposition 2.3, see also [50].

Proposition 2.4. *Let $\mathcal{A}\in\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$ be a real tensor. Then, $\lambda\mathbf{x}_1\otimes\cdots\otimes\mathbf{x}_d$ with $\mathbf{x}=(\mathbf{x}_1,\dots,\mathbf{x}_d)\in\mathbb{S}^{n_1-1}\times\cdots\times\mathbb{S}^{n_d-1}$ and $\lambda\geq 0$ is a best rank one approximation of $\mathcal{A}$ if and only if $\lambda=\langle\mathcal{A},\tau(\mathbf{x})\rangle$ and $\mathbf{x}$ is a singular vector tuple of $\mathcal{A}$ corresponding to the largest singular value.*

2.4. **Łojasiewicz's inequality.** Next, we recall Łojasiewicz's inequalities which will play a key role in later analysis. The classical Łojasiewicz inequality for analytic functions are given below (cf. [44]):

> **(Classical Łojasiewicz's gradient inequality)** If f is an analytic function with $f(\mathbf{0})=0$ and $\nabla f(\mathbf{0})=\mathbf{0}$, then there exist positive constants μ,κ, and ϵ such that
>
> $$\|\nabla f(\mathbf{x})\|\geq\mu|f(\mathbf{x})|^{\kappa}\ \text{ for all }\ \|\mathbf{x}\|\leq\epsilon.$$

As pointed out in [3, 5], it is often difficult to determine the corresponding exponent κ in Łojasiewicz's gradient inequality, and they typically stay unknown. Some estimates of the exponent κ in the gradient inequality were derived by D'Acunto and Kurdyka in [15] in the case when f is a polynomial. We will recall this fundamental result in the next lemma, which will play a key role in our sublinear convergence rate analysis.

Lemma 2.5. *(Łojasiewicz's Gradient Inequality for Polynomials* [15, Theorem 4.2] *) Let f be a real polynomial on $\mathbb{R}^n$ with degree $d\in\mathbb{N}$. Suppose that $f(\mathbf{0})=0$ and $\nabla f(\mathbf{0})=\mathbf{0}$. Then there exist constants $c,\epsilon>0$ such that for all $\|\mathbf{x}\|\leq\epsilon$ we have*

$$\|\nabla f(\mathbf{x})\|\geq c|f(\mathbf{x})|^{\kappa}\ \textit{ with }\ \kappa=1-\frac{1}{d(3d-3)^{n-1}}.$$

We note that the Łojasiewicz inequality has recently been extended to broad classes of nonsmooth functions see [3,5,41] and references therein. Moreover, the above explicit exponent estimates of Łojasiewicz's inequality for polynomials have also been extended to classes of nonsmooth semi-algebraic functions [39, 40]. These extensions have found important applications in nonsmooth optimization, stability analysis and spectral theory of tensors [39,40,42].

3. Sublinear convergence rate analysis for the higher order power method

In this section, we will provide an overall worst case analysis on the convergence rate of the classical higher order power method for finding the best rank one approximation of a real tensor. It is known that the whole sequence generated by this algorithm converges to a singular vector tuple [45,59], while the convergence rate issue is much more complicated than it is expected [24]. It can converge superlinearly, linearly or sublinearly. On the other hand, the explicit worst case (sublinear) convergence rate is still an important issue which is not completely understood [58,63]. Our results in this section will provide an answer for this question. We start by recalling the higher order power method.

3.1. **The higher order power method.** Let the ambient tensor space be $\mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$. We now study the iterative sequence generated by a higher order power method for the best rank one approximation of a given tensor. To avoid triviality, we will only consider nonzero tensors. For a given nonzero tensor $\mathcal{A} \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$, one can find a vector $\mathbf{x}^{(0)} = (\mathbf{x}_1^{(0)}, \dots, \mathbf{x}_d^{(0)}) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}$ such that $\langle \mathcal{A}, \tau(\mathbf{x}^{(0)}) \rangle \neq 0$. By normalizing if necessary, we may assume that $\mathbf{x}^{(0)} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$.

Let $F : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d} \to \mathbb{R} \cup \{\infty\}$ be defined as

$$F(\mathbf{x}) := -\langle \mathcal{A}, \tau(\mathbf{x}) \rangle + \sum_{i=1}^{d} \delta_{\mathbb{S}^{n_i-1}}(\mathbf{x}_i) \text{ for all } \mathbf{x} \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}, \tag{7}$$

where $\delta_{\mathbb{S}^{n_i-1}}$ is the indicator function of the unit sphere in $\mathbb{R}^{n_i}$ for $i = 1, \dots, d$ given by

$$\delta_{\mathbb{S}^{n_i-1}}(\mathbf{a}) = \begin{cases} 0 & \text{if} \quad \mathbf{a} \in \mathbb{R}^{n_i},\ \|\mathbf{a}\| = 1, \\ +\infty & \text{else.} \end{cases}$$

It is easy to see that F is the summation of a polynomial and an indicator function over a compact set, and so, is a proper and lower semicontinuous function. Moreover, we have

$$\begin{aligned} \min\{F(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}\} &\iff \max\{\langle \mathcal{A}, \tau(\mathbf{x}) \rangle \mid \mathbf{x} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}\} \\ &\iff \min\{\|\mathcal{A} - \lambda\tau(\mathbf{x})\|_{\text{HS}}^2 \mid \mathbf{x} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1} \text{ and } \lambda \in \mathbb{R}\}, \end{aligned} \tag{8}$$

of which the last is the best rank one approximation problem for the given tensor $\mathcal{A}$ (cf. Section 2.3). Note that the problem formulation in (8) is different from that in [58, 59], in which the objective function is $\frac{1}{2}\|\mathcal{A} - \mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_d\|^2$ and there is no normalizations on $\mathbf{x}_i$'s. Thus, if $(\mathbf{x}_1, \dots, \mathbf{x}_d)$ is a critical point, so is $(\gamma_1\mathbf{x}_1, \dots, \gamma_d\mathbf{x}_d)$ for any γ_i's such that $\prod_{i=1}^d \gamma_i = 1$. These make us considering (7) to possibly simplify the analysis on nondegenerate critical points.

A standard algorithm for solving (8) is the higher order power method (HOPM) in the literature [17, 18, 59], which can be regarded as a generalization of the classical power method for matrices to tensors.

Algorithm 3.1. *Higher Order Power Method: best rank one approximation of tensors*

Input a nonzero tensor $\mathcal{A}$.

Step 0 (Initialization): choose $\mathbf{x}^{(0)} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$ *such that* $\langle \mathcal{A}, \tau(\mathbf{x}^{(0)}) \rangle \neq 0$. *Let* $\mathbf{x}_{0,l} = \mathbf{x}_l^{(0)}$ *for* $l = 1, \dots, d$, $k := 1$ *and* $i := 1$.

Step 1: Let $\mathbf{u}^{(k,i)} := (\mathbf{x}_{k,1}, \dots, \mathbf{x}_{k,i-1}, \mathbf{0}, \mathbf{x}_{k-1,i+1}, \dots, \mathbf{x}_{k-1,d})$ *and*

$$\tau_i(\mathbf{u}^{(k,i)}) = \mathbf{x}_{k,1} \otimes \cdots \otimes \mathbf{x}_{k,i-1} \otimes \mathbf{x}_{k-1,i+1} \otimes \cdots \otimes \mathbf{x}_{k-1,d}.$$

Compute $\mathbf{x}_{k,i}$ *via*

$$\lambda_{k,i} := \|\mathcal{A}\tau_i(\mathbf{u}^{(k,i)})\| \text{ and } \mathbf{x}_{k,i} := \frac{\mathcal{A}\tau_i(\mathbf{u}^{(k,i)})}{\lambda_{k,i}}. \tag{9}$$

Form

$$\mathbf{x}^{(k,i)} := (\mathbf{x}_{k,1}, \dots, \mathbf{x}_{k,i}, \mathbf{x}_{k-1,i+1}, \dots, \mathbf{x}_{k-1,d}). \tag{10}$$

Step 2: If $i = d$, *then go to Step 3. Otherwise, let* $i := i + 1$ *and go back to Step 1.*

Step 3: Let $k := k + 1$ *and* $i := 1$, *go back to Step 1.*

Note that

$$\tau_i(\mathbf{x}^{(k,i)}) = \tau_i(\mathbf{u}^{(k,i)}) = \mathbf{x}_{k,1} \otimes \cdots \otimes \mathbf{x}_{k,i-1} \otimes \mathbf{x}_{k-1,i+1} \otimes \cdots \otimes \mathbf{x}_{k-1,d}.$$

The construction of Algorithm 3.1 implies that, for all $k = 1, 2, \dots$ and $i = 1, \dots, d$,

$$\lambda_{k,i} = \|\mathcal{A}\tau_i(\mathbf{x}^{(k,i)})\| \text{ and } \mathcal{A}\tau_i(\mathbf{x}^{(k,i)}) = \lambda_{k,i}\mathbf{x}_{k,i}. \tag{11}$$

Before proceeding, we make a convention on the notation. We let

$$\mathbf{x}^{(k)} := (\mathbf{x}_{k,1},\dots,\mathbf{x}_{k,d}) \text{ and } \mathbf{x}^{(k+1,0)} := \mathbf{x}^{(k)} \text{ for all } k = 0,1,\dots.$$

Thus,

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k,d)} = \mathbf{x}^{(k+1,0)} \text{ for all } k = 0,1,\dots,$$

and

$$\mathcal{A}\tau_{i+1}(\mathbf{x}^{(k,i)}) = \mathcal{A}\tau_{i+1}(\mathbf{x}^{(k,i+1)}) \text{ for all } k = 0,1,\dots, \text{ and } i = 0,\dots,d-1. \tag{12}$$

The global convergence of Algorithm 3.1 is known (cf. [59, 60]). We shall establish the explicit sublinear convergence rate of this algorithm through the next three subsections.

3.2. **Sufficient decrease.** In the following, we will first show that $\lambda_{k,i} > 0$ for all $k = 1,2,\dots,$ and $i = 1,\dots,d$, and hence Algorithm 3.1 is well-defined.

Note that $\tau_1(\mathbf{x}^{(1,1)}) = \tau_1(\mathbf{x}^{(0)}) = \mathbf{x}_2^{(0)} \otimes \cdots \otimes \mathbf{x}_d^{(0)}$. This together with $\langle \mathcal{A}, \tau(\mathbf{x}^{(0)})\rangle \neq 0$ implies that

$$\lambda_{1,1} = \|\mathcal{A}\tau_1(\mathbf{x}^{(1,1)})\| = \|\mathcal{A}\tau_1(\mathbf{x}^{(0)})\| > 0.$$

From the second relation of (11), one has, for all $k = 1,2,\dots,$ and $i = 1,\dots,d$,

$$\lambda_{k,i}\,\mathbf{x}_{k,i} = \mathcal{A}\tau_i(\mathbf{x}^{(k,i)}),$$

and so,

$$\lambda_{k,i} = \lambda_{k,i}\|\mathbf{x}_{k,i}\|^2 = \big(\mathcal{A}\tau_i(\mathbf{x}^{(k,i)})\big)^{\mathsf{T}}\mathbf{x}_{k,i} = \langle \mathcal{A}, \tau(\mathbf{x}^{(k,i)})\rangle. \tag{13}$$

Hence, it is easy to see that $\lambda_{k,i}$'s are obtained by alternatively minimizing $\mathbf{x}_i$'s in (7), and thus $\lambda_{k,i}$ is monotonically increasing

$$0 < \lambda_{1,1} \le \cdots \le \lambda_{1,d} \le \dots \le \lambda_{k,1} \le \dots \le \lambda_{k,d} \le \lambda_{k+1,1} \le \cdots \le \lambda_{k+1,d} \le \dots. \tag{14}$$

This shows that the denominator in the quotient of the second relation of (9) is positive, and thus Algorithm 3.1 is well-defined.

Define $\lambda_k = \lambda_{k,d}$, $k = 1,2,\dots$. For the sake of convention, we let

$$\lambda_{k+1,0} = \lambda_k \text{ for all } k = 1,2,\dots.$$

It is an immediate fact from Algorithm 3.1 that

$$\mathbf{x}^{(k)} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}, \tag{15}$$

which with Cauchy-Schwartz's inequality or a direct calculation implies that

$$\lambda_k \le \|\mathcal{A}\|_{\mathrm{HS}} \text{ for all } k = 1,2,\dots. \tag{16}$$

This and (14) give us that the sequence $\{\lambda_k\}$ monotonically converges.

It follows from the update of $\mathbf{x}^{(k,i)}$ that

$$\lambda_{k+1} - \lambda_k \ge \sum_{i=1}^{d} \frac{\lambda_{1,1}}{2}\|\mathbf{x}_{k+1,i} - \mathbf{x}_{k,i}\|^2 = \frac{\lambda_{1,1}}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \text{ for all } k = 1,2,\dots.$$

Then, it, together with $F(\mathbf{x}^{(k)}) = -\langle \mathcal{A}, \tau(\mathbf{x}^{(k)})\rangle = -\lambda_k$ and (15), implies that

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k+1)}) \ge \frac{\lambda_{1,1}}{2}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|^2 \text{ for all } k = 1,2,\dots. \tag{17}$$

3.3. **Relative error.** In the following, we will obtain a lower bound of the change of the values of the objective function F between the current iteration and the next iteration in terms of the gradient information of a function related to F.

Let

$$G(\mathbf{x}) := \langle \mathcal{A}, \tau(\mathbf{x}) \rangle. \tag{18}$$

It follows from the iteration that (cf. (11))

$$\langle \mathcal{A}, \tau(\mathbf{x}^{(k,i)}) \rangle \mathbf{x}_{k,i} = \mathcal{A}\tau_i(\mathbf{x}^{(k,i)}) \text{ for all } i = 1, \ldots, d \text{ and } k = 1, 2, \ldots.$$

Let $\nabla_i G(\mathbf{x}) \in \mathbb{R}^{n_i}$ denote the partial gradient of G with respect to the i-th block vector $\mathbf{x}_i$. Therefore, for all $i = 1, \ldots, d$, we have

$$\begin{aligned} \|\nabla_i G(\mathbf{x}^{(k)}) - \lambda_k \mathbf{x}_{k,i}\| &= \|\mathcal{A}\tau_i(\mathbf{x}^{(k)}) - \lambda_k \mathbf{x}_{k,i}\| \\ &= \|\mathcal{A}\tau_i(\mathbf{x}^{(k)}) - \mathcal{A}\tau_i(\mathbf{x}^{(k+1,i)}) + \lambda_{k+1,i}\mathbf{x}_{k+1,i} - \lambda_{k+1,i}\mathbf{x}_{k,i} + \lambda_{k+1,i}\mathbf{x}_{k,i} - \lambda_k \mathbf{x}_{k,i}\| \\ &\le \|\mathcal{A}\tau_i(\mathbf{x}^{(k)}) - \mathcal{A}\tau_i(\mathbf{x}^{(k+1,i)})\| + \|\lambda_{k+1,i}(\mathbf{x}_{k+1,i} - \mathbf{x}_{k,i})\| + \|(\lambda_{k+1,i} - \lambda_k)\mathbf{x}_{k,i}\| \\ &\le \|\mathcal{A}\|_{\mathrm{HS}} \|\tau_i(\mathbf{x}^{(k+1)}) - \tau_i(\mathbf{x}^{(k+1,i)})\| + \|\mathcal{A}\|_{\mathrm{HS}} \|\mathbf{x}_{k+1,i} - \mathbf{x}_{k,i}\| + |\lambda_{k+1,i} - \lambda_k|, \end{aligned}$$

where the second equality follows from the fact that $\mathcal{A}\tau_i(\mathbf{x}^{(k+1,i)}) = \lambda_{k+1,i}\mathbf{x}_{k+1,i}$, the first inequality is from the triangle inequality and the last inequality follows from (14), (16) and the fact that $\|\mathbf{x}_{k,i}\| = 1$. Thus, it follows from (15), $\lambda_{k+1,i} = \langle \mathcal{A}, \tau(\mathbf{x}^{(k+1,i)}) \rangle$, $\lambda_k = \langle \mathcal{A}, \tau(\mathbf{x}^{(k)}) \rangle$ and Lemma 2.1 that, for each $i = 1, \ldots, d$,

$$\|\nabla_i G(\mathbf{x}^{(k)}) - \lambda_k \mathbf{x}_{k,i}\| \le 2\sqrt{d}\|\mathcal{A}\|_{\mathrm{HS}} \|\mathbf{x}^{(k+1,i)} - \mathbf{x}^{(k+1)}\| + \|\mathcal{A}\|_{\mathrm{HS}} \|\mathbf{x}_{k+1,i} - \mathbf{x}_{k,i}\| \le L\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|,$$

where $L = (2\sqrt{d} + 1)\|\mathcal{A}\|_{\mathrm{HS}} > 0$. Therefore,

$$\|\nabla G(\mathbf{x}^{(k)}) - \lambda_k \mathbf{x}^{(k)}\| \le \sqrt{d}L\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|. \tag{19}$$

Combining (19) and (17), we have that

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k+1)}) \ge \frac{\lambda_{1,1}}{2dL^2} \|\nabla G(\mathbf{x}^{(k)}) - \lambda_k \mathbf{x}^{(k)}\|^2. \tag{20}$$

3.4. **Sublinear convergence rate analysis.** Since the iterative sequence $\{\mathbf{x}^{(k)}\}$ is in the compact set $\mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$, it has an accumulation point $\mathbf{x}^*$ which is also the limit of this sequence [59].

Next, we establish a sublinear convergence with an explicit sublinear convergence rate for the higher order power method. This complements the work of [59] by providing an explicit estimate of the sublinear convergence rate. Similar to [59], our approach is built on the tools of Łojasiewicz's inequality. On the other hand, different to [59], our analysis make use of a Lagrangian-type function of the function $F(\mathbf{x}) := -\langle \mathcal{A}, \tau(\mathbf{x}) \rangle + \sum_{i=1}^d \delta_{\mathbb{S}^{n_i-1}}(\mathbf{x}_i)$ given in (7) while [59] used the merit function $\tilde{G}(\mathbf{x}) := \frac{1}{2}\|\mathcal{A} - \tau(\mathbf{x})\|_{\mathrm{HS}}^2$. We also note that it is also possible to deduce sublinear convergence rate by using the merit function $\tilde{G}$ directly. But, as demonstrated in Remark 3.3, our approach here produces a much improved sublinear convergence rate estimate.

Theorem 3.2 (**Sublinear Convergence Rate of Higher Order Power Method).** *Let $\{\mathbf{x}^{(k)}\}$ be generated by Algorithm 3.1 for a given nonzero tensor $\mathcal{A} \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$. Let $N = n_1 + \cdots + n_d$ and $p = d(3d-3)^N$. Then, the following statements hold.*

(1) *Let $\sigma_k = \langle \mathcal{A}, \tau(\mathbf{x}^{(k)}) \rangle$. Then, $\{\sigma_k\}$ globally converges to a singular value of $\mathcal{A}$, denoted as σ^*, with sublinear convergence rate at least $O(k^{-\frac{p}{p-2}})$, that is, there exist $M_1 > 0$ and $k_0 \in \mathbb{N}$ such that for all $k \ge k_0$*

$$|\sigma_k - \sigma^*| \le M_1\, k^{-\frac{p}{p-2}}.$$

(2) $\{\mathbf{x}^{(k)}\}$ *converges to* $\mathbf{x}^*$ *globally with the sublinear convergence rate at least* $O(k^{-\frac{1}{p-2}})$*, that is, there exist* $M_2 > 0$ *and* $k_0 \in \mathbb{N}$ *such that for all* $k \geq k_0$

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq M_2\, k^{-\frac{1}{p-2}}.$$

Remark 3.3. Before we proceed to the proof, we note that, one can also obtain a sublinear convergence rate, by using Lemma 2.5 and the merit function suggested in [59]

$$\tilde{G}(\mathbf{x}) = \frac{1}{2}\|\mathcal{A} - \tau(\mathbf{x})\|_{\text{HS}}^2.$$

On the other hand, this produces a weaker sublinear convergence rate than the result presented in Theorem 3.2. Indeed, note that $\tilde{G}$ is a real polynomial with dimension $N = n_1 + \cdots + n_d$ and degree $2d$. Direct verification (by following the same method of proof in Theorem 3.2) shows that this approach leads to the result that the higher order power method generates a sequence $\mathbf{x}^{(k)} \to \mathbf{x}^*$ with a sublinear convergence rate $O(k^{-\frac{1}{2d(6d-3)^{N-1}-2}})$, whereas the rate of Theorem 3.2 is $O(k^{-\frac{1}{d(3d-3)^N-2}})$. Noting that, for all $d \geq 3$ (and so, $N = n_1 + \ldots + n_d \geq d \geq 3$),

$$2d(6d-3)^{N-1} - 2 > d(3d-3)^N - 2,$$

showing that the convergence rate established in Theorem 3.2 is stronger. Moreover, as $6d - 3 > 2(3d-3)$ and $N = n_1 + \ldots + n_d \geq d$,

$$\frac{2d(6d-3)^{N-1} - 2}{d(3d-3)^N - 2} \approx \frac{2^N}{3d-3} \geq \frac{2^d}{3d-3} \to \infty \text{ as } d \to \infty \text{ or } N \to \infty.$$

So, our current approach gives a significant better convergence rate estimate in the case d or N is large.

Proof. [Proof of (1)] As shown in [59], $\mathbf{x}^{(k)} \to \mathbf{x}^*$ for some $\mathbf{x}^* \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$. Actually, $\mathbf{x}^*$ is a singular vector tuple with singular value $\langle \mathcal{A}, \tau(\mathbf{x}^*)\rangle$ by [45] (see also [59]), i.e., $\mathcal{A}\tau_i(\mathbf{x}^*) = \langle \mathcal{A}, \tau(\mathbf{x}^*)\rangle \mathbf{x}_i^*$ for all $i \in \{1, \ldots, d\}$. Recall that $G(\mathbf{x}) = \langle \mathcal{A}, \tau(\mathbf{x})\rangle$ and define a function $H : (\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}) \times \mathbb{R} \to \mathbb{R}$ by

$$H(\mathbf{x}, \mu) = -G(\mathbf{x}) + \mu \sum_{i=1}^{d} (\|\mathbf{x}_i\|^2 - 1) = -\langle \mathcal{A}, \tau(\mathbf{x})\rangle + \mu \sum_{i=1}^{d} (\|\mathbf{x}_i\|^2 - 1).$$

Let $\mu^* = \frac{\langle \mathcal{A}, \tau(\mathbf{x}^*)\rangle}{2}$ and let $\widehat{H}(\mathbf{x}, \mu) := H(\mathbf{x}, \mu) - H(\mathbf{x}^*, \mu^*)$. Then, it is clear that $\widehat{H}$ is a real polynomial on $\mathbb{R}^{N+1}$ of degree d with $\widehat{H}(\mathbf{x}^*, \mu^*) = 0$ and $\nabla \widehat{H}(\mathbf{x}^*, \mu^*) = 0$. Actually, by the expression of $\widehat{H}(\mathbf{x}, \mu)$ and $H(\mathbf{x}, \mu)$, we then have that

$$\nabla \widehat{H}(\mathbf{x}^*, \mu^*) = \nabla H(\mathbf{x}^*, \mu^*) = (2\mu^* \mathbf{x}_1^* - \mathcal{A}\tau_1(\mathbf{x}^*), \ldots, 2\mu^* \mathbf{x}_d^* - \mathcal{A}\tau_d(\mathbf{x}^*)) = \mathbf{0}.$$

Applying Lemma 2.5, it then follows that there exist $c, \epsilon > 0$ such that

$$\|\nabla \widehat{H}(\mathbf{x}, \mu)\| \geq c\, |\widehat{H}(\mathbf{x}, \mu)|^\tau \text{ for all } \|(\mathbf{x}, \mu) - (\mathbf{x}^*, \mu^*)\| \leq \epsilon,$$

where $\tau = 1 - \left(d(3d-3)^N\right)^{-1}$. Then, for all $\mathbf{x} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$ and $\mu \in \mathbb{R}$ with $\|(\mathbf{x}, \mu) - (\mathbf{x}^*, \mu^*)\| \leq \epsilon$, one has

$$\| - \nabla G(\mathbf{x}) + 2\mu \mathbf{x}\|^2 \geq c^2 \left(\langle \mathcal{A}, \tau(\mathbf{x}^*)\rangle - \langle \mathcal{A}, \tau(\mathbf{x})\rangle\right)^{2\tau}.$$

Now, note that $|\langle \mathcal{A}, \tau(\mathbf{x})\rangle - \langle \mathcal{A}, \tau(\mathbf{x}^*)\rangle| \leq \sqrt{d}\, \|\mathcal{A}\|_{\text{HS}} \|\mathbf{x} - \mathbf{x}^*\|$ by Lemma 2.1. By choosing a smaller ϵ if necessary, it follows by letting $\mu = \frac{\langle \mathcal{A}, \tau(\mathbf{x})\rangle}{2}$ that for all $\mathbf{x} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$ with $\|\mathbf{x} - \mathbf{x}^*\| \leq \epsilon$

$$\| - \nabla G(\mathbf{x}) + \langle \mathcal{A}, \tau(\mathbf{x})\rangle \mathbf{x}\|^2 \geq c^2 \left(\langle \mathcal{A}, \tau(\mathbf{x}^*)\rangle - \langle \mathcal{A}, \tau(\mathbf{x})\rangle\right)^{2\tau} = c^2 \left(F(\mathbf{x}) - F(\mathbf{x}^*)\right)^{2\tau}. \tag{21}$$

Recall that $F(\mathbf{x}) := -\langle \mathcal{A}, \tau(\mathbf{x})\rangle + \sum_{i=1}^d \delta_{\mathbb{S}^{n_i-1}}(\mathbf{x}_i)$ given in (7). From (20), we have

$$\begin{aligned} F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k+1)}) &\geq \frac{\lambda_{1,1}}{2dL^2}\|\nabla G(\mathbf{x}^{(k)}) - \lambda_k \mathbf{x}^{(k)}\|^2 \\ &= \frac{\lambda_{1,1}}{2dL^2}\|\nabla G(\mathbf{x}^{(k)}) - \langle \mathcal{A}, \tau(\mathbf{x}^{(k)})\rangle \mathbf{x}^{(k)}\|^2, \end{aligned}$$

where the last equality follows from $\lambda_k = \langle \mathcal{A}, \tau(\mathbf{x}^{(k)})\rangle$. As $\mathbf{x}^{(k)} \to \mathbf{x}^*$, there exists k_0 such that for all $k \geq k_0$, $\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \epsilon$. So, this together with (21) implies that, for all $k \geq k_0$,

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k+1)}) \geq \frac{\lambda_{1,1}c^2}{2dL^2}\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big)^{2\tau}.$$

Denote $\beta_k = F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*) \geq 0$ and $M = \frac{\lambda_{1,1}c^2}{2dL^2}$. Then, $\beta_k \geq \beta_{k+1} + M\beta_k^{2\tau}$. Let $h(x) = x^{-2\tau}$. It follows that

$$\beta_k - \beta_{k+1} \geq M\beta_k^{2\tau} = Mh(\beta_k)^{-1}.$$

Noting that $h(x) = f'(x)$ with $f(x) = \frac{x^{1-2\tau}}{1-2\tau}$ and h is non-increasing on $\mathbb{R}_{++}$, it follows that

$$M \leq h(\beta_k)(\beta_k - \beta_{k+1}) \leq \int_{\beta_{k+1}}^{\beta_k} h(x)dx = f(\beta_k) - f(\beta_{k+1}) = \frac{1}{1-2\tau}\big(\beta_k^{1-2\tau} - \beta_{k+1}^{1-2\tau}\big) = \frac{1}{2\tau-1}\big(\beta_{k+1}^{1-2\tau} - \beta_k^{1-2\tau}\big).$$

So,

$$\beta_k^{1-2\tau} \geq M(2\tau-1) + \beta_{k-1}^{1-2\tau} \geq \cdots \geq M(2\tau-1)(k-k_0) + \beta_{k_0}^{1-2\tau},$$

and hence, there exists $C > 0$ such that for all $k \geq k_0$,

$$0 \leq \beta_k \leq C\, k^{-\frac{1}{2\tau-1}}.$$

Now, note that $\beta_k = F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*) = -\langle \mathcal{A}, \tau(\mathbf{x}^{(k)})\rangle + \langle \mathcal{A}, \tau(\mathbf{x}^*)\rangle$ and $\frac{1}{2\tau-1} = \frac{1}{1-2p^{-1}} = \frac{p}{p-2}$. So, for all $k \geq k_0$,

$$|\sigma_k - \sigma^*| \leq C\, k^{-\frac{p}{p-2}}.$$

Thus, statement of (i) follows.

[Proof of (2)] From (19), one has

$$\|\nabla G(\mathbf{x}^{(k)}) - \lambda_k \mathbf{x}^{(k)}\| \leq \sqrt{d}L\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|, \tag{22}$$

and so, (21) gives us that

$$\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big)^{\tau} \leq \frac{\sqrt{d}L}{c}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|.$$

Let $s_k = \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|$. Then, we have

$$\beta_k^{\tau} \leq \frac{\sqrt{d}L}{c}s_k. \tag{23}$$

Moreover, noting that $\varphi : s \mapsto -s^{1-\tau}$ is convex on $\mathbb{R}_{++}$, one has

$$\varphi(\beta_{k+1}) - \varphi(\beta_k) \geq \varphi'(\beta_k)(\beta_{k+1} - \beta_k).$$

This gives us that

$$\beta_k^{1-\tau} - \beta_{k+1}^{1-\tau} \geq (1-\tau)\beta_k^{-\tau}\big(\beta_k - \beta_{k+1}\big) = (1-\tau)\beta_k^{-\tau}\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k+1)})\big).$$

This together with (17) shows that

$$\beta_k^{1-\tau} - \beta_{k+1}^{1-\tau} \geq (1-\tau)\frac{\lambda_{1,1}}{2}\beta_k^{-\tau}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|^2 = (1-\tau)\frac{\lambda_{1,1}}{2}\beta_k^{-\tau}s_k^2. \tag{24}$$

Therefore, this, together with (23), implies that there exists $M > 0$ such that

$$s_k \leq M\big(\beta_k^{1-\tau} - \beta_{k+1}^{1-\tau}\big). \tag{25}$$

For any $N > N_0 > k_0$, summing this from N_0 to N gives that

$$\sum_{k=N_0}^{N} s_k \leq M\big(\beta_{N_0}^{1-\tau} - \beta_{N+1}^{1-\tau}\big).$$

Letting $N \to \infty$ and note that $\beta_N \to 0$ as $N \to \infty$, we see that

$$\sum_{k=N_0}^{\infty} s_k \leq M\beta_{N_0}^{1-\tau}.$$

This shows that $\sum_{k=1}^{\infty} s_k < +\infty$. Denote $\Delta_k = \sum_{i=k}^{\infty} s_i$, $k \geq k_0$. Then, from (23), one has

$$\Delta_k \leq M\left(\frac{\sqrt{d}L}{c} s_k\right)^{\frac{1-\tau}{\tau}}.$$

As $\frac{1-\tau}{\tau} < 1$, there exists $C > 0$ such that

$$\Delta_k^{\frac{\tau}{1-\tau}} \leq Cs_k = C\left(\Delta_k - \Delta_{k+1}\right).$$

In other words, one has $\Delta_k \geq \Delta_{k+1} + \frac{1}{C}\Delta_k^{\frac{\tau}{1-\tau}}$. Then, a similar line method as in part one shows that there exists $C_0 > 0$ such that

$$\Delta_k \leq C_0 k^{-\frac{1-\tau}{2\tau-1}}.$$

Finally, the conclusion follows by noting that $\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \sum_{i=k}^{\infty} \|\mathbf{x}^{(i)} - \mathbf{x}^{(i+1)}\| = \Delta_k$. □

Remark 3.4. **(Discussion on the convergence rate estimate)** In Theorem 3.2, the convergence rate of the objective function value sequence $\{\langle \mathcal{A}, \tau(\mathbf{x}^{(k)})\rangle\}$ is $O(k^{-\frac{p}{p-2}})$ with $1 < \frac{p}{p-2} < 2$. The convergence rate estimate we obtained here can be regarded as "suboptimal" from optimization point of view. To see this, we recall that HOPM can be regarded as a special instance of the block coordinate descent method (see the introduction), and it was demonstrated that the objective function value of block coordinate descent method (or sometimes referred as alternating minimization methods) at least exhibits a sublinear convergence rate of $O(\frac{1}{k})$ without assuming strong convexity assumptions (cf. [4, Theorem 14.11]). Moreover, it is also widely known from the celebrated work of Nesterov that the optimal rate for first-order methods (which includes the block coordinate descent method) is $O(1/k^2)$ even in the convex setting [47, Section 2.2], and the optimal rate usually can only be achieved by incorporating additional information and generating the next iterate from both the previous iterate and the current iterate (known as Nesterov acceleration techniques). Therefore, one can view our convergence rate estimate as "suboptimal" from optimization point of view.

Finally, note that $p = d(3d-3)^N$. Our derived sublinear convergence estimate indeed depends on the dimension N and the order of the underlying tensor space d. In particular, when N or d increase, the quantity $k^{-\frac{p}{p-2}}$ and $k^{-\frac{1}{p-2}}$ decrease. As a result, when N and d increase (and so, the problem becomes more complex), the derived convergence rate becomes slower.

Remark 3.5. We note from (14), (16) and $\lambda_k = \lambda_{k,d} = \langle \mathcal{A}, \tau(\mathbf{x}^{(k)}\rangle$ that

$$0 < \lambda_k \leq \lambda_{k+1} \leq \|\mathcal{A}\|_{\text{HS}} \text{ and } \lambda_k \to \langle \mathcal{A}, \tau(\mathbf{x}^*)\rangle > 0.$$

This shows that the HOPM indeed converges to a singular vector tuple with positive singular value.

Next, we will examine when the higher order power method will exhibit a linear convergence rate. To achieve this, we will introduce and examine the notion of nondegenerate singular vector tuples in the next section. This notion plays an important role in establishing the desired linear convergence rate of higher order power method later on.

4. Nondegenerate singular vector tuples and linear convergence of higher order power method

In this section, we introduce the notion of nondegenerate singular vector tuples and show that it serves as a sufficient condition ensuring the linear convergence of higher order power method. Our result does not require the limit point to be a local/global solution of the corresponding optimization formulation, and so, distinguishes with the sufficient conditions for linear convergence in the literature (e.g. [58, 63]). More importantly, we will further justify, in the next section, that for almost all tensors (in the sense of Lebesgue measure), each of its singular vector tuples is nondegenerate. As a result, the HOPM *exhibits global R-linear convergence rate* for almost all tensors with any initialization.

We start by recalling some standard facts on smooth manifold. We refer to [7, 21] for basic notions of functions on a smooth manifold. Let $M \subseteq \mathbb{R}^n$ be a manifold and let $f : M \to \mathbb{R}$ be a smooth function. Let $\mathbf{x} \in M$ and let $\mathrm{T}_{\mathbf{x}}(M)$ be the tangent space of M at $\mathbf{x}$. Denote the geodesic curve at this point in a tangent direction $\mathbf{u} \in \mathrm{T}_{\mathbf{x}}(M)$ by $\mathbf{x}(t)$. The manifold *gradient* of f on the manifold M at $\mathbf{x} \in M$ is denoted as $\mathrm{grad}(f)(\mathbf{x})$, and is determined by $\frac{d}{dt}f(\mathbf{x}(t))|_{t=0}$. Similarly, the manifold *Hessian* on the manifold M at $\mathbf{x} \in M$ is denoted as $\mathrm{Hess}(f)(\mathbf{x})$, and is determined by $\frac{d^2}{dt^2}f(\mathbf{x}(t))|_{t=0}$. We say a point $\mathbf{x} \in M$ a *nondegenerate critical point* of the function f if the manifold gradient satisfies $\mathrm{grad}(f)(\mathbf{x}) = \mathbf{0}$ and the manifold Hessian $\mathrm{Hess}(f)(\mathbf{x})$ is a nonsingular linear operator from the tangent space $\mathrm{T}_{\mathbf{x}}(M)$ to itself. We will also need the following lemma, which is the well-known Morse's theorem [6, 46].

Lemma 4.1. *Let M be a smooth manifold with $\dim(M) = k \le n$ and $f : M \subseteq \mathbb{R}^n \to \mathbb{R}$ be a smooth function for which $\mathbf{x}^*$ is a nondegenerate critical point. Then, in some neighborhood $\mathcal{U}$ of $\mathbf{x}^*$ in M, there is a local C^∞ coordinate system, namely a C^∞ diffeomorphism*

$$\varphi : \mathcal{U} \to \mathcal{V} \subset \mathbb{R}^k$$

with $\varphi(\mathbf{x}^*) = \mathbf{0}$ and a neighborhood $\mathcal{V}$ of $\mathbf{0}$ such that the map $\tilde{f} = f \circ \varphi^{-1}$ takes the form $\tilde{f}(\mathbf{x}) = f(\mathbf{x}^*) - x_1^2 - \ldots - x_s^2 + x_{s+1}^2 + \ldots + x_k^2$ for some nonnegative integer s.

The following proposition shows that a form of Łojasiewicz inequality with exponent 1/2 can be achieved under the nondegeneracy assumption on the critical points.

Proposition 4.2. *Let M be a smooth manifold with $\dim(M) = k \le n$ and $f : M \subseteq \mathbb{R}^n \to \mathbb{R}$ be a smooth function for which $\mathbf{x}^*$ is a nondegenerate critical point. Then there exists a neighborhood $\mathcal{U}$ in M of $\mathbf{x}^*$ such that for all $\mathbf{x} \in \mathcal{U}$*

$$\| \mathrm{grad}(f)(\mathbf{x})\|^2 \ge \kappa |f(\mathbf{x}) - f(\mathbf{x}^*)|$$

for some $\kappa > 0$.

Proof. Let us adopt the notation in Lemma 4.1. In the new coordinate system, locally, we have

$$\begin{aligned}
|f(\varphi^{-1}(\mathbf{y})) - f(\mathbf{x}^*)| = |\tilde{f}(\mathbf{y}) - \tilde{f}(\mathbf{0})| &= |y_1^2 + \cdots + y_s^2 - y_{s+1}^2 - \cdots - y_k^2| \\
&\le \|\mathbf{y}\|^2 \\
&= \frac{1}{4}\|\nabla \tilde{f}(\mathbf{y})\|^2 \\
&= \frac{1}{4}\left\| \begin{bmatrix} \langle \mathrm{grad}(f)(\mathbf{x}), \mathrm{D}(\varphi^{-1})(\mathbf{y})\xi_1\rangle \\ \vdots \\ \langle \mathrm{grad}(f)(\mathbf{x}), \mathrm{D}(\varphi^{-1})(\mathbf{y})\xi_k\rangle \end{bmatrix} \right\|^2 \\
&\le \frac{1}{4}\| \mathrm{D}(\varphi^{-1})(\mathbf{y})\|^2 \| \mathrm{grad}(f)(\mathbf{x})\|^2,
\end{aligned}$$

where $\varphi(\mathbf{x}) = \mathbf{y}$, $\mathrm{D}(\varphi^{-1})(\mathbf{y})$ is the differential of φ^{-1} which is a linear mapping from $\mathbb{R}^k$ to the tangent space $\mathrm{T}_{\mathbf{x}}(M)$, and $\xi_i \in \mathbb{R}^k$ is the i-th basis vector of $\mathbb{R}^k$. Since φ is a coordinate system, $\mathrm{D}(\varphi^{-1})(\mathbf{y}) = [\mathrm{D}(\varphi)(\mathbf{x})]^{-1}$ is nonsingular. Thus, $\|\mathrm{D}(\varphi^{-1})(\mathbf{y})\|$ is bounded in a small neighborhood of $\mathbf{x}^*$. Consequently, the result follows. □

Definition 4.3 (Nondegenerate Singular Vector Tuples). *We say a singular vector tuple of $\mathcal{A}$ is nondegenerate if $\mathbf{x}$ is a nondegenerate critical point of G on $\mathbb{S}$.*

Next, we show that the HOPM exhibits R-linear convergence rate when the limit point $\mathbf{x}^*$ is a nondegenerate singular vector tuple of $\mathcal{A}$. Note that in the special case where $\mathbf{x}^*$ is a local minimizer of $G(x) = \langle \mathcal{A}, \tau(\mathbf{x})\rangle$, the next Theorem 4.4 reduces to the linear convergence result established in [58] (similar conclusion was achieved in the cases where $\mathbf{x}^*$ is a global minimizer in [63, Theorem 4.3]). On the other hand, our next Theorem 4.4 can be applied to the cases where the limit point $\mathbf{x}^*$ is not a local minima (for example, when it is a saddle point of G). More importantly, we will justify, in the next section, that our assumption "the limit point $\mathbf{x}^*$ is a nondegenerate singular vector tuple" is satisfied for almost all tensors.

Theorem 4.4 (Linear Convergence of HOPM under Nondegeneracy). *Let $\{\mathbf{x}^{(k)}\}$ be generated by Algorithm 3.1 for a given nonzero tensor $\mathcal{A}$ with $\mathbf{x}^{(k)} \to \mathbf{x}^*$. Suppose that $\mathbf{x}^*$ is a nondegenerate singular vector tuple of $\mathcal{A}$. Then, the sequence $\{\mathbf{x}^{(k)}\}$ converges to $\mathbf{x}^*$ globally with R-linear convergence rate, that is, there exist $M > 0$, $r \in (0,1)$ and $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,*

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq M\, r^k.$$

Proof. Let $G(x) = \langle \mathcal{A}, \tau(\mathbf{x})\rangle$ be given as in (18) and let $F(\mathbf{x}) = -\langle \mathcal{A}, \tau(\mathbf{x})\rangle + \sum_{i=1}^d \delta_{\mathbb{S}^{n_i-1}}(\mathbf{x}_i)$ given in (7). It follows from Definition 4.3 that $\mathbf{x}^*$ is a nondegenerate critical point of G on the joint sphere. Then, this together with Proposition 4.2 implies that there exists $\kappa > 0$ such that

$$\|\operatorname{grad}(G)(\mathbf{x})\|^2 \geq \kappa |G(\mathbf{x}^*) - G(\mathbf{x})| \tag{26}$$

for all $\mathbf{x} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$ sufficiently close to $\mathbf{x}^*$.

From [22, Section 2.4], we see that the manifold gradient of G at a point $\mathbf{x} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$ is

$$\operatorname{grad}(G)(\mathbf{x}) = \nabla G(\mathbf{x}) - \operatorname{diag}\{\nabla_{\mathbf{x}_1} G(\mathbf{x})^{\mathsf{T}}\mathbf{x}_1 I, \dots, \nabla_{\mathbf{x}_d} G(\mathbf{x})^{\mathsf{T}}\mathbf{x}_d I\}\mathbf{x}, \tag{27}$$

This implies that

$$\begin{aligned} \operatorname{grad}(G)(\mathbf{x}^{(k+1)}) &= \nabla G(\mathbf{x}^{(k+1)}) - \operatorname{diag}\{\nabla_{\mathbf{x}_1} G(\mathbf{x}^{(k+1)})^{\mathsf{T}}\mathbf{x}_1^{(k+1)} I, \dots, \nabla_{\mathbf{x}_d} G(\mathbf{x}^{(k+1)})^{\mathsf{T}}\mathbf{x}_d^{(k+1)} I\}\mathbf{x}^{(k+1)} \\ &= \nabla G(\mathbf{x}^{(k+1)}) - \lambda_{k+1}\mathbf{x}^{(k+1)}, \end{aligned} \tag{28}$$

where we used the fact that $\nabla_{\mathbf{x}_i} G(\mathbf{x}^{(k+1)})^{\mathsf{T}}\mathbf{x}_i^{(k+1)} = \langle \mathcal{A}\tau_i(\mathbf{x}^{(k+1)}), \mathbf{x}_i^{(k+1)}\rangle = \langle \mathcal{A}, \tau(\mathbf{x}^{(k+1)})\rangle = \lambda_{k+1}$.

Let $\lambda^* = \langle \mathcal{A}, \tau(\mathbf{x}^*)\rangle = G(\mathbf{x}^*) = -F(\mathbf{x}^*)$. Then, as established in equations (16) and (17), $\lambda_k = -F(\mathbf{x}^{(k)})$ monotonically increases and converges to λ^*. This, together with $\mathbf{x}^{(k+1)} \in \mathbb{S}^{n_1-1} \times \cdots \times \mathbb{S}^{n_d-1}$ and (7), implies that

$$G(\mathbf{x}^*) - G(\mathbf{x}^{(k+1)}) = F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) \geq 0. \tag{29}$$

On the other hand, we have

$$\begin{aligned} \|\nabla_i G(\mathbf{x}^{(k+1)}) - \lambda_{k+1}\mathbf{x}_{k+1,i}\| &= \|\mathcal{A}\tau_i(\mathbf{x}^{(k+1)}) - \lambda_{k+1}\mathbf{x}_{k+1,i}\| \\ &= \|\mathcal{A}\tau_i(\mathbf{x}^{(k+1)}) - \mathcal{A}\tau_i(\mathbf{x}^{(k+1,i)}) + \mathcal{A}\tau_i(\mathbf{x}^{(k+1,i)}) - \lambda_{k+1}\mathbf{x}_{k+1,i}\| \\ &\leq \|\mathcal{A}\tau_i(\mathbf{x}^{(k+1)}) - \mathcal{A}\tau_i(\mathbf{x}^{(k+1,i)})\| + \|(\lambda_{k+1,i} - \lambda_{k+1})\mathbf{x}_{k+1,i}\| \\ &\leq \|\mathcal{A}\|_{\mathrm{HS}}\|\tau_i(\mathbf{x}^{(k+1)}) - \tau_i(\mathbf{x}^{(k+1,i)})\| + |\lambda_{k+1,i} - \lambda_{k+1}|. \end{aligned}$$

Similar to (19), we have

$$\|\nabla G(\mathbf{x}^{(k+1)}) - \lambda_{k+1}\mathbf{x}^{(k+1)}\| \leq \sqrt{d}L\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|,$$

where $L = (2\sqrt{d}+1)\|\mathcal{A}\|_{\mathrm{HS}}$.

Since $\mathbf{x}^{(k)} \to \mathbf{x}^*$, there exists $k_0 > 0$ such that (26) holds for all $k \geq k_0$. Thus, for all $k \geq k_0$ we have

$$\begin{aligned} F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k+1)}) &\geq \frac{\lambda_{1,1}}{2dL^2}\|\nabla G(\mathbf{x}^{(k+1)}) - \lambda_{k+1}\mathbf{x}^{(k+1)}\|^2 \\ &= \frac{\lambda_{1,1}}{2dL^2}\|\operatorname{grad}(G)(\mathbf{x}^{(k+1)})\|^2 \\ &\geq \frac{\lambda_{1,1}\kappa}{2dL^2}\big(F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*)\big), \end{aligned}$$

where the first inequality follows from preceding estimation and (17), the equality is from (28), and the last inequality from (26) and (29). Thus, for all $k \geq k_0$

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) \leq \frac{2dL^2}{2dL^2 + \lambda_{1,1}\kappa}\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big).$$

It also follows from (17) that for all $k \geq k_0$

$$\begin{aligned} \|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\| &\leq \sqrt{\frac{2}{\lambda_{1,1}}}\sqrt{F(\mathbf{x}^{(i)}) - F(\mathbf{x}^{(i+1)})} \\ &\leq \sqrt{\frac{2}{\lambda_{1,1}}}\sqrt{F(\mathbf{x}^{(i)}) - F(\mathbf{x}^*)} \\ &\leq \sqrt{\frac{2}{\lambda_{1,1}}}\left[\sqrt{\frac{2dL^2}{2dL^2 + \lambda_{1,1}\kappa}}\right]^{i-1}\sqrt{F(\mathbf{x}^{(1)}) - F(\mathbf{x}^*)}. \end{aligned}$$

Thus, we have $\sum_{i=k}^{\infty}\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\| < +\infty$. Since $\mathbf{x}^{(k)} \to \mathbf{x}^*$,

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \sum_{i=k}^{\infty}\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\|.$$

This implies that for all $k \geq k_0$

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \sqrt{\frac{2}{\lambda_{1,1}}}\sqrt{F(\mathbf{x}^{(1)}) - F(\mathbf{x}^*)}\,\frac{1}{1 - \sqrt{\frac{2dL^2}{2dL^2+\lambda_{1,1}\kappa}}}\left[\sqrt{\frac{2dL^2}{2dL^2 + \lambda_{1,1}\kappa}}\right]^{k-1},$$

which is the claimed R-linear convergence. Thus, the conclusion follows. □

5. Typical linear convergence of higher order power method

In this section, we establish that the HOPM *exhibits global R-linear convergence rate* for almost all tensors. We achieve this by showing that, for almost all tensors, each of its singular vector tuples is nondegenerate. This gives an answer to the "linear convergence" behavior of HOPM mentioned in the literature, such as in [60, Page 1603]: "*We shall not concern ourselves with the rate of convergence, though it is expected to be linear.*"

To do this, we will need the following notion of Morse function and a theorem regarding its existence. Let M be a smooth manifold, a function $f : M \to \mathbb{R}$ is called a *Morse function* if each critical point of f on M is nondegenerate. The following result on existence of Morse functions is well-known, see for example [7, Proposition 17.18].

Lemma 5.1. **(Morse function is "typical")** *Let M be a manifold of dimension m in $\mathbb{R}^n$. For almost all $\mathbf{a} := (a_1, \dots, a_n)^\mathsf{T} \in \mathbb{R}^n$, the function*

$$f(\mathbf{x}) = a_1 x_1 + \cdots + a_n x_n$$

is a Morse function on M.

We will also need the following proposition on critical points of functions over two diffeomorphic smooth manifolds. Recall that two smooth manifolds M_1 and M_2 are called locally diffeomorphic if there is a mapping $\phi : M_1 \to M_2$ such that for each point $\mathbf{x} \in M_1$ there exist a neighborhood $U \subseteq M_1$ of $\mathbf{x}$ and a neighborhood $V \subseteq M_2$ of $\phi(\mathbf{x})$ so that the restriction mapping $\phi : U \to V$ is a diffeomorphism [21]. In this case, the corresponding ϕ is called a local diffeomorphism between M_1 and M_2.

Proposition 5.2. *Let $M_1 \subseteq \mathbb{R}^{n_1}$ and $M_2 \subset \mathbb{R}^{n_2}$ be two diffeomorphic smooth manifolds of the same dimension $m \le \min\{n_1, n_2\}$ and let $\phi : M_1 \to M_2$ be the corresponding local diffeomorphism. Let $f : M_2 \to \mathbb{R}$ be a smooth function. Then $\mathbf{x} \in M_1$ is a (nondegenerate) critical point of $f \circ \phi$ on M_1 if and only if $\phi(\mathbf{x})$ is a (nondegenerate) critical point of f on M_2.*

Proof. Let a local coordinate system of M_1 in a neighborhood of $\mathbf{x}$ be $\alpha_1, \dots, \alpha_m : U_1 \subseteq M_1 \to V_1 \subseteq \mathbb{R}^m$. Here U_1 is a neighborhood of $\mathbf{x}$ and V_1 a neighborhood of $\mathbf{p} \in \mathbb{R}^m$ such that

$$(\alpha_1(\mathbf{x}), \dots, \alpha_m(\mathbf{x})) = \mathbf{p}.$$

The inverse of $\alpha : U_1 \to V_1$ is denoted as β. Similarly, let $\theta : M_2 \to \mathbb{R}^m$ be a local coordinate system of M_2 in a neighborhood of $\phi(\mathbf{x})$, and η be the inverse of θ with $\theta(\phi(\mathbf{x})) = \mathbf{q}$ for some $\mathbf{q} \in \mathbb{R}^m$. The corresponding neighborhoods are respectively U_2, V_2. As ϕ being a local diffeomorphism, we can shrink the neighborhoods if necessary to satisfy $U_2 = \phi(U_1)$. In the following, we always work in these neighborhoods without tedious repetitions.

Let $\mathbf{t} = \alpha(\mathbf{x})$ be the local coordinates of M_1, and $\mathbf{s} = \theta(\phi(\mathbf{x}))$ be the local coordinates of M_2. We then have a relation between the two local coordinates

$$\mathbf{t} = \alpha \circ \phi^{-1} \circ \eta(\mathbf{s}). \tag{30}$$

Note that

$$d\alpha_i \in \mathrm{T}_{\mathbf{x}} M_1,\ d\phi^{-1} : \mathrm{T}_{\phi(\mathbf{x})} M_2 \to \mathrm{T}_{\mathbf{x}} M_1,\ \text{and } d\eta : \mathbb{R}^m \to \mathrm{T}_{\phi(\mathbf{x})} M_2$$

in the neighborhood of $\mathbf{x}$ (as well as $\phi(\mathbf{x})$). We then have

$$\frac{\partial t_j}{\partial s_i} = d\alpha_j d\phi^{-1} d\eta(\mathbf{e}_i),$$

where $\mathbf{e}_i$ is the i-th coordinate vector in $\mathbb{R}^m$. Let the matrix $H \in \mathbb{R}^{m \times m}$ be defined as

$$h_{ij} = \frac{\partial t_j}{\partial s_i}(\mathbf{q}).$$

So, as a linear operator,

$$H = d\alpha\, d\phi^{-1}\, d\eta.$$

Since all of α, ϕ and η are local diffeomorphisms, we see that H is a nonsingular linear operator.

With the above preparation, we have the identity

$$(f \circ \phi)(\mathbf{x}) = (f \circ \phi \circ \beta)(\mathbf{t}) = (f \circ \phi \circ \beta \circ \alpha \circ \phi^{-1} \circ \eta)(\mathbf{s}) = (f \circ \eta)(\mathbf{s}).$$

Since a point on a manifold being a (nondegenerate) critical point of a smooth function is independent of the choices of coordinates, it follows that $\mathbf{x}$ is a critical point of $f \circ \phi$ on M_1 if and only if

$$\frac{\partial (f \circ \phi \circ \beta)(\mathbf{t})}{\partial t_i}(\mathbf{p}) = 0 \text{ for all } i \in \{1, \dots, m\}$$

as $\beta(\mathbf{p}) = \mathbf{x}$. Likewise, $\phi(\mathbf{x})$ is a critical point of f on M_2 if and only if

$$\frac{\partial(f\circ\eta)(\mathbf{s})}{\partial s_i}(\mathbf{q}) = 0 \text{ for all } i \in \{1,\dots,m\}$$

as $\eta(\mathbf{q}) = \phi(\mathbf{x})$.

However, (cf. (30))

$$\begin{aligned}\frac{\partial(f\circ\eta)(\mathbf{s})}{\partial s_i}(\mathbf{q}) &= \frac{\partial(f\circ\phi\circ\beta\circ\alpha\circ\phi^{-1}\circ\eta)(\mathbf{s})}{\partial s_i}(\mathbf{q})\\ &= \sum_{j=1}^m \frac{\partial(f\circ\phi\circ\beta)(\mathbf{t})}{\partial t_j}(\mathbf{p})\frac{\partial t_j}{\partial s_i}(\mathbf{q}) \text{ for all } i\in\{1,\dots,m\} \quad (31)\end{aligned}$$

since $\mathbf{p} = \alpha\circ\phi^{-1}\circ\eta(\mathbf{q})$.

Let the vectors $\mathbf{a}\in\mathbb{R}^m$ and $\mathbf{b}\in\mathbb{R}^m$ be defined as

$$\mathbf{a} = \left(\frac{\partial(f\circ\phi\circ\beta)(\mathbf{t})}{\partial t_1}(\mathbf{p}),\dots,\frac{\partial(f\circ\phi\circ\beta)(\mathbf{t})}{\partial t_m}(\mathbf{p})\right)^{\mathsf{T}},$$

and

$$\mathbf{b} = \left(\frac{\partial(f\circ\eta)(\mathbf{s})}{\partial s_1}(\mathbf{q}),\dots,\frac{\partial(f\circ\eta)(\mathbf{s})}{\partial s_m}(\mathbf{q})\right)^{\mathsf{T}}.$$

Then, it follows from (31) that

$$\mathbf{b} = H\mathbf{a}.$$

As H being nonsingular, we thus conclude that $\mathbf{x}$ is a critical point of $f\circ\phi$ on M_1 if and only if $\phi(\mathbf{x})$ is a critical point of f on M_2.

In the following, we assume that $\mathbf{x}$ is a critical point of $f\circ\phi$ on M_1, i.e., $\mathbf{a}=\mathbf{0}$. Equivalently, $\mathbf{b}=\mathbf{0}$.

Let the matrix $A\in\mathbb{R}^{m\times m}$ be defined as

$$a_{ij} =: \frac{\partial^2(f\circ\phi\circ\beta)(\mathbf{t})}{\partial t_i\partial t_j}(\mathbf{p}) \text{ for all } i,j\in\{1,\dots,m\}.$$

Likewise, let the matrix $B\in\mathbb{R}^{m\times m}$ be defined as

$$b_{ij} =: \frac{\partial^2(f\circ\eta)(\mathbf{s})}{\partial s_i\partial s_j}(\mathbf{q}) \text{ for all } i,j\in\{1,\dots,m\}.$$

We have

$$\begin{aligned} b_{ij} &= \sum_{l=1}^m\sum_{k=1}^m \frac{\partial^2(f\circ\phi\circ\beta)(\mathbf{t})}{\partial t_k\partial t_l}(\mathbf{p})\frac{\partial t_k}{\partial s_i}(\mathbf{q})\frac{\partial t_l}{\partial s_j}(\mathbf{q}) + \sum_{k=1}^m \frac{\partial(f\circ\phi\circ\beta)(\mathbf{t})}{\partial t_k}(\mathbf{p})\frac{\partial^2 t_k}{\partial s_i\partial s_j}(\mathbf{q})\\ &= \sum_{l=1}^m\sum_{k=1}^m \frac{\partial^2(f\circ\phi\circ\beta)(\mathbf{t})}{\partial t_k\partial t_l}(\mathbf{p})\frac{\partial t_k}{\partial s_i}(\mathbf{q})\frac{\partial t_l}{\partial s_j}(\mathbf{q}),\end{aligned}$$

where the equality follows from $\mathbf{a}=\mathbf{0}$. Thus,

$$B = HAH^{\mathsf{T}}.$$

As H being nonsingular, the matrix A is nonsingular if and only if B is nonsingular. In other words, $\mathbf{x}$ is a nondegenerate critical point of $f\circ\phi$ on M_1 if and only if $\phi(\mathbf{x})$ is a nondegenerate critical point of f on M_2, since degeneracy is independent of the choices of local coordinates. □

Now, we are in the position to present one of our main results, showing that the function G is a Morse function on the joint sphere for almost all tensors.

Theorem 5.3 (Singular Vector Tuples are "typically" nondegenerate). *For almost all tensors in $\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$, each of its singular vector tuples is nondegenerate.*

Proof. Define

$$M=\{\mathcal{A}\in\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}\mid \mathcal{A}=\tau(\mathbf{x})\text{ with }\mathbf{x}\in\mathbb{S}=\mathbb{S}^{n_1-1}\times\cdots\times\mathbb{S}^{n_d-1}\}$$

be the image of the Segre mapping restricted on the joint sphere $\mathbb{S}$ (cf. (1)).

We claim that M is a smooth manifold which is *locally diffeomorphic* to the joint sphere $\mathbb{S}$, i.e., there is a mapping $\phi:\mathbb{S}\to M$ such that for each $\mathbf{x}\in\mathbb{S}$, there exists a neighborhood U of $\mathbf{x}$ in $\mathbb{S}$ such that the restriction $\phi:U\to V$ is a diffeomorphism for a neighborhood V of M. Granting this, to conclude the proof, it suffices to show that for almost all tensors $\mathcal{A}\in\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$, the function

$$G(\mathbf{x})=\langle\mathcal{A},\tau(\mathbf{x})\rangle$$

is a Morse function over the joint sphere $\mathbb{S}$. Let $\psi:\mathbb{S}\to M$ be the Segre mapping from the joint sphere to the manifold M, then we have

$$G(\mathbf{x})=(\tilde{G}\circ\psi)(\mathbf{x}),$$

where $\tilde{G}(\mathcal{U})=\langle\mathcal{A},\mathcal{U}\rangle$ for all $\mathcal{U}\in\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$. From Lemma 5.1, we see that $\tilde{G}$ is a Morse function over the manifold M for almost all $\mathcal{A}\in\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$. Note that ψ is a surjective local diffeomorphism from $\mathbb{S}$ to M. Thus, the conclusion follows from Proposition 5.2.

We now justify our claim that M is a smooth manifold which is *locally diffeomorphic* to the joint sphere $\mathbb{S}$. Given a point $\mathbf{x}=(\mathbf{x}_1,\dots,\mathbf{x}_d)\in\mathbb{S}$, let

$$i_j\in\operatorname{argmax}\{|(\mathbf{x}_i)_j|\mid j\in\{1,\dots,n_i\}\}\text{ for all }i\in\{1,\dots,d\}.$$

Obviously,

$$|(\mathbf{x}_i)_{i_j}|\geq\frac{1}{\sqrt{n_i}}\text{ for all }i\in\{1,\dots,d\}.$$

If we take $\epsilon<\frac{1}{2}\min\{\frac{1}{\sqrt{n_i}}\mid i\in\{1,\dots,n\}\}$, then for every

$$\mathbf{y}\in\mathbb{S}\cap\{\mathbf{w}\mid\|\mathbf{w}-\mathbf{x}\|\leq\epsilon\},\tag{32}$$

we have

$$\operatorname{sign}([\psi(\mathbf{y})]_{i_1\dots i_d})=\operatorname{sign}([\psi(\mathbf{x})]_{i_1\dots i_d}),\tag{33}$$

since $(\mathbf{y}_i)_{i_j}$'s have constant sign over the neighborhood (32). In the following, we will show that $\psi:U:=\mathbb{S}\cap\{\mathbf{w}\mid\|\mathbf{w}-\mathbf{x}\|\leq\epsilon\}\to M$ is a local diffeomorphism from U to $V:=\psi(U)$. To see this, let

$$T(i_1\dots i_d):=\{\mathcal{A}\in\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}\mid a_{i_1\dots i_d}\neq 0\}.$$

Then, we have a smooth mapping $\phi_{i_1\dots i_d}:V\subset T(i_1\dots i_d)\cap M\to U\subset\mathbb{R}^{n_1}\times\cdots\times\mathbb{R}^{n_d}$ from V to U as

$$\phi_{i_1\dots i_d}(\mathcal{A})=(\mathbf{y}_1,\dots,\mathbf{y}_d)$$

where

$$\mathbf{y}_1:=\kappa_1\frac{(a_{1i_2\dots i_d},\dots,a_{n_1i_2\dots i_d})^{\mathsf{T}}}{\|(a_{1i_2\dots i_d},\dots,a_{n_1i_2\dots i_d})\|},\dots,\mathbf{y}_d:=\kappa_d\frac{(a_{i_1i_2\dots 1},\dots,a_{i_1i_2\dots n_d})^{\mathsf{T}}}{\|(a_{i_1i_2\dots 1},\dots,a_{i_1i_2\dots n_d})\|},\tag{34}$$

and $\kappa_i\in\{-1,1\}$ is a constant for all $i\in\{1,\dots,d\}$ such that

$$\phi_{i_1\dots i_d}(\psi(\mathbf{x}))=(\mathbf{x}_1,\dots,\mathbf{x}_d).$$

It follows from (33) and (34) that $\phi_{i_1\dots i_d}\circ\psi$ is the identity over U. Moreover, direct verification shows that $\psi\circ\phi_{i_1\dots i_d}$ equals the identity mapping over U. So, we see that ψ is a local diffeomorphism. It is well-known that the joint sphere is a smooth manifold. Thus, M is a smooth manifold which is locally diffeomorphic to the joint sphere which is of dimension $\sum_{i=1}^{d}n_i-d$. $\square$

Corollary 5.4 (Linear convergence of HOPM is "typical"). *Let $\{\mathbf{x}^{(k)}\}$ be generated by Algorithm 3.1 with $\mathbf{x}^{(k)} \to \mathbf{x}^*$. Then, for almost all $\mathcal{A} \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$, the sequence $\{\mathbf{x}^{(k)}\}$ converges to $\mathbf{x}^*$ globally with R-linear convergence rate.*

Corollary 5.4 shows that, for almost all tensors, the HOPM exhibits global R-linear convergence rate regardless the initialization. It is also interesting and worth identifying a class of tensors such that HOPM converges R-linearly for *each tensor* in this class . We will show that the orthogonally decomposable tensors form such a class.

6. Orthogonally decomposable tensors

In this section, we show that, without any regularity assumptions, HOPM always converges R-linearly for orthogonally decomposable tensors with order at least 3. We shall achieve this by establishing the fact that every nonzero singular vector tuple of an orthogonally decomposable tensor with order at least 3 is nondegenerate. To establish our desired result, we will further study nondegenerate singular vector tuples.

Note that the nondegeneracy is defined via the nonsingularity of the manifold Hessian of G on the joint sphere $\mathbb{S}$. Moreover, the manifold Hessian of G is defined as a linear operator from the tangent space $\mathrm{T}_{\mathbf{x}}(\mathbb{S})$ to itself. As the tangent space depends on the critical point $\mathbf{x}$, it may not be so convenient to analyze manifold Hessian. Therefore, to obtain further insight of the nondegenerate singular vector tuples, we will introduce a square system of equations motivated from the defining equations of singular vector tuples.

To that end, we first rewrite the system of equations for singular vectors of a tensor $\mathcal{A}$ into a square system with respect to the variable vector $\mathbf{x}$ as

$$S(\mathbf{x}) := \begin{cases} \mathcal{A}\tau_1(\mathbf{x}) - \langle \mathcal{A}, \tau(\mathbf{x})\rangle \mathbf{x}_1 = \mathbf{0}, \\ \qquad \vdots \\ \mathcal{A}\tau_d(\mathbf{x}) - \langle \mathcal{A}, \tau(\mathbf{x})\rangle \mathbf{x}_d = \mathbf{0}. \end{cases} \tag{35}$$

Here, we emphasize that, unlike in the definition of singular vector tuples, the spherical normalization is not required explicitly any more. On the other hand, it is easy to see that whenever $\langle \mathcal{A}, \tau(\mathbf{x})\rangle \neq 0$, we necessarily have that

$$\mathbf{x}_i \in \mathbb{S}^{n_i-1} = \{\mathbf{y} \in \mathbb{R}^{n_i} \mid \|\mathbf{y}\| = 1\} \text{ for all } i = 1, \ldots, d.$$

However, when $\langle \mathcal{A}, \tau(\mathbf{x})\rangle = 0$, this *automatic normalization* for $\mathbf{x}_i$'s becomes vacant. In particular, there exists trivial solutions $\mathbf{x} \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}$ with at least two block vectors being zeros and the rest being arbitrary.

For simplicity, we will abbreviate a singular vector tuple corresponding to a nonzero singular value as a *nonzero singular vector tuple*.

Given a continuously differentiable mapping $g : \mathbb{R}^n \to \mathbb{R}^n$, recall that a solution of the equation system $g(\mathbf{x}) = \mathbf{0}$ is said to be *nonsingular* if the Jacobian matrix $\nabla g(\mathbf{x})$ is nonsingular. The next proposition shows that any nonsingular solution of $S(\mathbf{x}) = \mathbf{0}$ associates with a nonzero singular value, and so, is indeed a nonzero singular vector tuple.

Proposition 6.1 (Nonsingular solutions of (35) associate with nonzero singular value)**.** *Let $(\mathbf{x}_1, \ldots, \mathbf{x}_d)$ be a nonsingular solution of $S(\mathbf{x}) = \mathbf{0}$ with $\sigma = \langle \mathcal{A}, \tau(\mathbf{x})\rangle$. Then $\sigma \neq 0$.*

Proof. First note that

$$\sigma = \langle \mathcal{A}, \tau(\mathbf{x})\rangle.$$

It is easy to calculate the Jacobian matrix of (35) as

$$\begin{aligned}\nabla_{\mathbf{x}} S(\mathbf{x}) &= -\sigma\,\mathrm{diag}\{I, I, \dots, I\}\\ &+ \begin{bmatrix} -\sigma\mathbf{x}_1\mathbf{x}_1^{\mathsf{T}} & \mathcal{A}\tau_{1,2}(\mathbf{x}) - \sigma\mathbf{x}_1\mathbf{x}_2^{\mathsf{T}} & \dots & \mathcal{A}\tau_{1,d}(\mathbf{x}) - \sigma\mathbf{x}_1\mathbf{x}_r^{\mathsf{T}} \\ \mathcal{A}\tau_{2,1}(\mathbf{x}) - \sigma\mathbf{x}_2\mathbf{x}_1^{\mathsf{T}} & -\sigma\mathbf{x}_2\mathbf{x}_2^{\mathsf{T}} & \dots & \mathcal{A}\tau_{2,d}(\mathbf{x}) - \sigma\mathbf{x}_2\mathbf{x}_d^{\mathsf{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{A}\tau_{d,1}(\mathbf{x}) - \sigma\mathbf{x}_d\mathbf{x}_1^{\mathsf{T}} & \mathcal{A}\tau_{d,2}(\mathbf{x}) - \sigma\mathbf{x}_d\mathbf{x}_2^{\mathsf{T}} & \dots & -\sigma\mathbf{x}_d\mathbf{x}_d^{\mathsf{T}} \end{bmatrix}\\ &= -\,\sigma\,\mathrm{diag}\{I, I, \dots, I\} - \sigma \begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots \\ \mathbf{x}_d\end{bmatrix} \begin{bmatrix}\mathbf{x}_1^{\mathsf{T}} & \mathbf{x}_2^{\mathsf{T}} & \dots & \mathbf{x}_d^{\mathsf{T}}\end{bmatrix} + \begin{bmatrix} 0 & \mathcal{A}\tau_{1,2}(\mathbf{x}) & \dots & \mathcal{A}\tau_{1,d}(\mathbf{x}) \\ \mathcal{A}\tau_{2,1}(\mathbf{x}) & 0 & \dots & \mathcal{A}\tau_{2,d}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{A}\tau_{d,1}(\mathbf{x}) & \mathcal{A}\tau_{d,2}(\mathbf{x}) & \dots & 0 \end{bmatrix}. \end{aligned} \tag{36}$$

If $\sigma = 0$, then

$$\nabla_{\mathbf{x}} S(\mathbf{x}) = \begin{bmatrix} 0 & \mathcal{A}\tau_{1,2}(\mathbf{x}) & \dots & \mathcal{A}\tau_{1,d}(\mathbf{x}) \\ \mathcal{A}\tau_{2,1}(\mathbf{x}) & 0 & \dots & \mathcal{A}\tau_{2,d}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{A}\tau_{d,1}(\mathbf{x}) & \mathcal{A}\tau_{d,2}(\mathbf{x}) & \dots & 0 \end{bmatrix}.$$

It is a direct calculation to see that

$$\nabla_{\mathbf{x}} S(\mathbf{x})\mathbf{x} = (d-1)\begin{bmatrix}\mathcal{A}\tau_1(\mathbf{x})\\ \vdots \\ \mathcal{A}\tau_d(\mathbf{x})\end{bmatrix} = (d-1)\sigma\mathbf{x} = \mathbf{0},$$

which contradicts the nonsingularity hypothesis. □

Next, we see that nonzero nondegenerate singular vector tuple can be characterized by nonsingular solutions of the algebraic equations $S(\mathbf{x}) = \mathbf{0}$ (cf (36)).

Proposition 6.2 (Characterizing nonzero nondegenerate singular vector tuples via nonsingular solutions of (36))**.** *Given a tensor $\mathcal{A} \in \mathbb{R}^{n_1} \otimes \dots \otimes \mathbb{R}^{n_d}$, a nonzero singular vector tuple $\mathbf{x}$ is nondegenerate if and only if $\mathbf{x}$ is a nonsingular solution of $S(\mathbf{x}) = \mathbf{0}$.*

Proof. A direct calculation shows that the manifold Hessian $\mathrm{Hess}(G)(\mathbf{x})$ of G at a singular vector tuple $\mathbf{x}$ is given by the formula

$$\langle \Delta^{(1)}, \mathrm{Hess}(G)(\mathbf{x})\Delta^{(2)}\rangle = \left\langle \Delta^{(1)}, \begin{bmatrix} -\sigma I & \mathcal{A}\tau_{1,2}(\mathbf{x}) & \dots & \mathcal{A}\tau_{1,d}(\mathbf{x}) \\ \mathcal{A}\tau_{2,1}(\mathbf{x}) & -\sigma I & \dots & \mathcal{A}\tau_{2,d}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{A}\tau_{d,1}(\mathbf{x}) & \mathcal{A}\tau_{d,2}(\mathbf{x}) & \dots & -\sigma I \end{bmatrix} \Delta^{(2)} \right\rangle, \tag{37}$$

for any two tangent vectors $\Delta^{(1)}, \Delta^{(2)} \in \mathrm{T}_{\mathbf{x}_1}(\mathbb{S}^{n_1-1}) \times \dots \times \mathrm{T}_{\mathbf{x}_d}(\mathbb{S}^{n_d-1})$, and in where $\sigma = G(\mathbf{x})$. For each $i = 1, \dots, d$, let $\mathrm{St}(n_i-1, n_i)$ be the Stiefel manifold consisting of $n_i \times (n_i - 1)$ matrices with orthonormal columns, and $P_i \in \mathrm{St}(n_i-1, n_i)$ be such that $P_i^{\mathsf{T}}\mathbf{x}_i = \mathbf{0}$. Then the columns of P_i form a basis for the tangent space $\mathrm{T}_{\mathbf{x}_i}(\mathbb{S}^{n_i-1})$. Therefore, $\mathrm{Hess}(G)(\mathbf{x})$ is singular if and only if there exist nonzero $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_d) \in \mathbb{R}^{n_1-1} \times \dots \times \mathbb{R}^{n_d-1}$ and $\alpha_i \in \mathbb{R}$ such that

$$\mathrm{Hess}(G)(\mathbf{x})\,\mathrm{diag}\{P_1, \dots, P_d\}\mathbf{y} = \left(\alpha_1\mathbf{x}_1^{\mathsf{T}}, \dots, \alpha_d\mathbf{x}_d^{\mathsf{T}}\right)^{\mathsf{T}}.$$

Note that the manifold Hessian of G is a linear operator from the tangent space $\mathrm{T}_{\mathbf{x}_1}(\mathbb{S}^{n_1-1}) \times \dots \times \mathrm{T}_{\mathbf{x}_d}(\mathbb{S}^{n_d-1})$ of the product manifold to itself [22]. So, $\mathrm{Hess}(G)(\mathbf{x})$ is singular if and only if there exist nonzero $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_d) \in \mathbb{R}^{n_1-1} \times \dots \times \mathbb{R}^{n_d-1}$ such that

$$\mathrm{Hess}(G)(\mathbf{x})\,\mathrm{diag}\{P_1, \dots, P_d\}\mathbf{y} = \mathbf{0}.$$

Now, suppose that $\mathbf{x}$ is a nonsingular solution. Then the matrix $\nabla_{\mathbf{x}} S(\mathbf{x})$ in (36) is nonsingular, and so, for any nonzero $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_d) \in \mathbb{R}^{n_1-1} \times \cdots \times \mathbb{R}^{n_d-1}$, we have

$$\mathbf{0} \neq \nabla_{\mathbf{x}} S(\mathbf{x}) \operatorname{diag}\{P_1, \ldots, P_d\}\mathbf{y} = \operatorname{Hess}(G)(\mathbf{x}) \operatorname{diag}\{P_1, \ldots, P_d\}\mathbf{y}.$$

This implies that the Hessian $\operatorname{Hess}(G)(\mathbf{x})$ is a nonsingular linear operator from the tangent space to itself. Thus, $\mathbf{x}$ is a nondegenerate singular vector tuple.

Conversely, suppose that $\operatorname{Hess}(G)(\mathbf{x})$ is a nonsingular linear operator from the tangent space to itself. We proceed by the method of contradiction and assume that $\nabla_{\mathbf{x}} S(\mathbf{x})$ is a singular matrix, i.e., $\nabla_{\mathbf{x}} S(\mathbf{x})\mathbf{z} = \mathbf{0}$ for some $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_d) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_d}$ with $\mathbf{z} \neq \mathbf{0}$. Write $\mathbf{z}_i = \alpha_i \mathbf{x}_i + \beta_i \mathbf{u}_i$, $i = 1, \ldots, d$ where $\alpha_i, \beta_i \in \mathbb{R}$ and $\mathbf{u}_i^{\mathsf{T}} \mathbf{x}_i = 0$ for all $i = 1, \ldots, d$. Then, we have

$$\nabla_{\mathbf{x}} S(\mathbf{x}) \begin{bmatrix} \alpha_1 \mathbf{x}_1 + \beta_1 \mathbf{u}_1 \\ \vdots \\ \alpha_d \mathbf{x}_d + \beta_d \mathbf{u}_d \end{bmatrix} = \mathbf{0}.$$

On the other hand, a direct calculation shows that

$$\nabla_{\mathbf{x}} S(\mathbf{x}) \begin{bmatrix} \alpha_1 \mathbf{x}_1 + \beta_1 \mathbf{u}_1 \\ \vdots \\ \alpha_d \mathbf{x}_d + \beta_d \mathbf{u}_d \end{bmatrix} = -2\sigma \begin{bmatrix} \alpha_1 \mathbf{x}_1 \\ \vdots \\ \alpha_d \mathbf{x}_d \end{bmatrix} + \operatorname{Hess}(G)(\mathbf{x}) \begin{bmatrix} \beta_1 \mathbf{u}_1 \\ \vdots \\ \beta_d \mathbf{u}_d \end{bmatrix}.$$

Since $\operatorname{Hess}(G)(\mathbf{x})$ maps a tangent vector into the tangent space, if the singular value $\sigma \neq 0$, we must have both

$$\begin{bmatrix} \alpha_1 \mathbf{x}_1 \\ \vdots \\ \alpha_d \mathbf{x}_d \end{bmatrix} = \mathbf{0} \text{ and } \operatorname{Hess}(G)(\mathbf{x}) \begin{bmatrix} \beta_1 \mathbf{u}_1 \\ \vdots \\ \beta_d \mathbf{u}_d \end{bmatrix} = \mathbf{0}.$$

This, together with the nonsingularity of $\operatorname{Hess}(G)(\mathbf{x})$, implies that $\beta_i \mathbf{u}_i = \gamma_i \mathbf{x}_i$ for some $\gamma_i \in \mathbb{R}$. It follows that $\beta_i \mathbf{u}_i = \mathbf{0}$ because if $\mathbf{x}_i \neq \mathbf{0}$, then $\gamma_i = \frac{\beta_i \mathbf{u}_i^T \mathbf{x}_i}{\|\mathbf{x}_i\|^2} = 0$ and so, $\beta_i \mathbf{u}_i = \mathbf{0}$. Thus, $\alpha_i \mathbf{x}_i = \beta_i \mathbf{u}_i = \mathbf{0}$ for all $i \in \{1, \ldots, d\}$. This contradicts the fact that $\mathbf{z} \neq \mathbf{0}$, and so, the conclusion follows. □

6.1. **Orthogonally decomposable tensors.** In this subsection, we show that the higher order power method always exhibits linear convergence when the underlying tensor is orthogonally decomposable and with order at least 3. To do this, recall that a tensor $\mathcal{A} \in \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d}$ is called *orthogonally decomposable* (cf. [34, 52, 63] [2]) if there exist orthogonal matrices

$$A_i = [\mathbf{a}_{i,1}, \ldots, \mathbf{a}_{i,n_i}] \in \mathbb{R}^{n_i \times n_i} \text{ for all } i = 1, \ldots, d$$

and numbers $\lambda_i \in \mathbb{R}$ for $i = 1, \ldots, D_0 := \min\{n_1, \ldots, n_d\}$ such that

$$\mathcal{A} = \sum_{i=1}^{D_0} \lambda_i \mathbf{a}_{1,i} \otimes \cdots \otimes \mathbf{a}_{d,i}. \tag{38}$$

Without loss of generality, we can assume that $\lambda_i \geq 0$ for all $i = 1, \ldots, D_0$. Note that some of the λ_i's can be zeros. By eliminating the zeros, we can further assume that a nonzero orthogonally decomposable tensor takes the form

$$\mathcal{A} = \sum_{i=1}^{D} \lambda_i \mathbf{a}_{1,i} \otimes \cdots \otimes \mathbf{a}_{d,i}.$$

[2] In [34], this notion was referred as completely orthogonally decomposable tensors.

with $D \leq D_0 := \min\{n_1,\dots,n_d\}$ and $\lambda_i > 0$, $i = 1,\dots,D$. It is worth noting that, for an orthogonally decomposable tensor with order $d \geq 3$, [63] established that its orthogonal decomposition is always unique (up to scaling). This uniqueness feature is not shared by the matrix cases. Throughout this subsection, we will always assume $d \geq 3$.

In the following, we will characterize all the nonzero singular vector tuples of an orthogonally decomposable tensor. As a preparation, we need the following observation of multilinear orthogonal invariance of (nondegenerate) singular vector tuples. Let $\mathbb{O}(n_i) \subset \mathbb{R}^{n_i\times n_i}$ be the group of orthogonal matrices and $Q_i \in \mathbb{O}(n_i)$ for all $i \in \{1,\dots,d\}$. For a given $\mathcal{A} \in \mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$ with $\mathcal{A} = (a_{j_1\dots j_d})$, the matrix-tensor multiplication $(Q_1,\dots,Q_d)\cdot\mathcal{A}$ is defined as a tensor in $\mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$ with its component being

$$\big[(Q_1,\dots,Q_d)\cdot\mathcal{A}\big]_{i_1\dots i_d} = \sum_{j_1=1}^{n_1}\cdots\sum_{j_d=1}^{n_d}(Q_1)_{i_1j_1}\cdots(Q_d)_{i_dj_d}a_{j_1\dots j_d} \quad \text{for all } i_j \in \{1,\dots,n_j\} \text{ and } j \in \{1,\dots,d\}.$$

Similar to matrices setting, one has the following *multilinear orthogonal invariance* property for singular vectors of tensors.

Proposition 6.3 (Multilinear Orthogonal Invariance)**.** *Let $\mathcal{A} \in \mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$ and $Q_i \in \mathbb{O}(n_i)$ for all $i \in \{1,\dots,d\}$. Then, $(\mathbf{x}_1,\dots,\mathbf{x}_d)$ is a singular vector tuple (respectively nondegenerate singular vector tuple) of $\mathcal{A}$ if and only if $(Q_1\mathbf{x}_1,\dots,Q_d\mathbf{x}_d)$ is a singular vector tuple (respectively nondegenerate singular vector tuple) of $(Q_1,\dots,Q_d)\cdot\mathcal{A}$.*

Proof. The results follow from a direct calculation, in particular noticing that the manifold Hessian $\mathrm{Hess}(G)(\mathbf{y})$ (cf. (37)) for the tensor $\mathcal{B} := (Q_1,\dots,Q_d)\cdot\mathcal{A}$ at the singular vector tuple $(\mathbf{y}_1,\dots,\mathbf{y}_d) = (Q_1\mathbf{x}_1,\dots,Q_d\mathbf{x}_d)$ is

$$\mathrm{diag}\{Q_1^\mathsf{T},\dots,Q_d^\mathsf{T}\}\,\mathrm{Hess}(G)(\mathbf{x})\,\mathrm{diag}\{Q_1,\dots,Q_d\}.$$

□

If $\mathcal{A}$ is an orthogonally decomposable tensor, then there exist $Q_i \in \mathbb{O}(n_i)$ for all $i \in \{1,\dots,d\}$ such that $(Q_1,\dots,Q_d)\cdot\mathcal{A}$ is a diagonal tensor with the diagonal elements being λ_i's and 0's. It follows from Proposition 6.3 that it is sufficient to study diagonal tensors for orthogonally decomposable tensors. Since HOPM finds nonzero singular vector tuples (and hence we consider only nonzero nondegenerate singular vector tuples), and the sub-vectors indexed by the zero diagonals of the singular vectors corresponding to a nonzero singular value should be zero, it is without loss of generality to assume that $D \leq n_1 \leq \cdots \leq n_d$ and $\mathcal{A} \in \mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$ is an d-th order diagonal tensor with the diagonal elements being positive λ_i's. In other words,

$$\mathcal{A} = \sum_{i=1}^{D}\lambda_i\mathbf{e}_i^1\otimes\cdots\otimes\mathbf{e}_i^d. \tag{39}$$

where $\lambda_i > 0$, $i = 1,\dots,D$, and, for each $i = 1,\dots,d$ and $j = 1,\dots,D$, $\mathbf{e}_j^i$ is the unit vector in $\mathbb{R}^{n_i}$ whose j-th coordinate is one and the other coordinates are zero.

Recall that the singular vector tuples enjoys the following symmetric property: $(\mathbf{x}_1,\dots,\mathbf{x}_d)$ is a singular vector tuple with singular value σ, if and only if $(\epsilon_1\mathbf{x}_1,\dots,\epsilon_d\mathbf{x}_d)$ is a singular vector tuple with singular value $-\sigma$ for any choices of $\epsilon_i \in \{-1,1\}$ with $\prod_{i=1}^d\epsilon_i = -1$. The next proposition says that nondegeneracy also enjoys a similar invariant property.

Proposition 6.4. *Let $\mathcal{A} \in \mathbb{R}^{n_1}\otimes\cdots\otimes\mathbb{R}^{n_d}$. Then, $\mathbf{x} = (\mathbf{x}_1,\dots,\mathbf{x}_d)$ is a nondegenerate singular vector tuple of $\mathcal{A}$ with singular value σ if and only if $\mathbf{y} = (\epsilon_1\mathbf{x}_1,\dots,\epsilon_d\mathbf{x}_d)$ is a nondegenerate singular vector tuple of $\mathcal{A}$ with singular value $(\prod_{i=1}^d\epsilon_i)\sigma$ for any choices $\epsilon_i \in \{-1,1\}$ for all $i \in \{1,\dots,d\}$.*

Proof. First of all, for any choices $\epsilon_i \in \{-1, 1\}$ for all $i \in \{1, \dots, d\}$, $\mathbf{y} = (\epsilon_1\mathbf{x}_1, \dots, \epsilon_d\mathbf{x}_d)$ is a singular vector tuple is clear. It is also clear that $\mathrm{T}_{\mathbf{y}_i}(\mathbb{S}^{n_i-1}) = \mathrm{T}_{\mathbf{x}_i}(\mathbb{S}^{n_i-1})$ for all $i \in \{1, \dots, d\}$. A direct calculation will see that

$$\begin{bmatrix} -(\Pi_{i=1}^d\epsilon_i)\sigma I & \mathcal{A}\tau_{1,2}(\mathbf{y}) & \dots & \mathcal{A}\tau_{1,d}(\mathbf{y}) \\ \mathcal{A}\tau_{2,1}(\mathbf{y}) & -(\Pi_{i=1}^d\epsilon_i)\sigma I & \dots & \mathcal{A}\tau_{2,d}(\mathbf{y}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{A}\tau_{d,1}(\mathbf{y}) & \mathcal{A}\tau_{d,2}(\mathbf{y}) & \dots & -(\Pi_{i=1}^d\epsilon_i)\sigma I \end{bmatrix} = (\Pi_{i=1}^d\epsilon_i)D \begin{bmatrix} -\sigma I & \mathcal{A}\tau_{1,2}(\mathbf{x}) & \dots & \mathcal{A}\tau_{1,d}(\mathbf{x}) \\ \mathcal{A}\tau_{2,1}(\mathbf{x}) & -\sigma I & \dots & \mathcal{A}\tau_{2,d}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{A}\tau_{d,1}(\mathbf{x}) & \mathcal{A}\tau_{d,2}(\mathbf{x}) & \dots & -\sigma I \end{bmatrix} D,$$

where $D = \mathrm{diag}\{\epsilon_1 I, \dots, \epsilon_d I\}$ is a diagonal matrix. Since each $\mathrm{T}_{\mathbf{y}_i}(\mathbb{S}^{n_i-1})$ is a linear space, we see that $D = \mathrm{diag}\{\epsilon_1 I, \dots, \epsilon_d I\}$ is a nonsingular linear mapping from the tangent space to itself. It then follows from (37) that the manifold Hessian of G is nonsingular at $\mathbf{y}$ if and only if it is nonsingular at $\mathbf{x}$. Thus, the conclusion follows. ☐

To see the nondegeneracy of all the nonzero singular vector tuples, from the preceding Proposition, we only need to consider the singular vector tuples with positive singular values. Using the algebraic description of singular vector tuples (cf. (4)), we see that the singular vector tuples with positive singular values of $\mathcal{A}$ in the form of (39) are: for all $k = 1, \dots, D$,

$$(\sigma, \mathbf{x}_1, \dots, \mathbf{x}_d) = \left(\left(\frac{1}{\sum_{i\in\Lambda_k}\lambda_i^{\frac{-2}{d-2}}}\right)^{\frac{d-2}{2}}, P_1\mathbf{w}, \dots, P_d\mathbf{w}\right), \tag{40}$$

where $\Lambda_k \subseteq \{1, \dots, D\}$ is a subset of cardinality $k \le D$,

$$\mathbf{w} = \left(\frac{1}{\sum_{i\in\Lambda_k}\lambda_i^{\frac{-2}{d-2}}}\right)^{\frac{1}{2}} (\lambda_1^{\frac{-1}{d-2}}, \dots, \lambda_D^{\frac{-1}{d-2}})^{\mathsf{T}},$$

and the matrices P_i satisfy the following property (P):

(P) $P_i \in \mathbb{R}^{n_i\times D}$ is a diagonal matrix with the (j, j)-th diagonal element satisfying

$$\begin{cases} (P_i)_{jj} \in \{-1, 1\}, & \text{if } j \in \Lambda_k, \\ (P_i)_{jj} = 0, & \text{otherwise,} \end{cases}$$

and, $\prod_{i=1}^d (P_i)_{jj} = 1$ for all $j \in \Lambda_k$.

Let $G_n := \left\{(\epsilon_1, \dots, \epsilon_n) \in \{-1, 1\}^n \mid \prod_{i=1}^n \epsilon_i = 1\right\}$. Define a multiplication on G_n by

$$g_1g_2 = (\epsilon_1\eta_1, \dots, \epsilon_n\eta_n) \text{ for all } g_1 = (\epsilon_1, \dots, \epsilon_n) \in G_n \text{ and } g_2 = (\eta_1, \dots, \eta_n) \in G_n.$$

It is a routine to verify that G_n is an abelian finite group under this multiplication and the order $\#(G_n)$ of G_n is 2^{n-1}. To count the total number of positive singular vector tuples, we shall first study a group action on $\mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_d}$ by G_d defined as

$$g \cdot \mathbf{x} = (\epsilon_1\mathbf{x}_1, \dots, \epsilon_d\mathbf{x}_d) \text{ for all } g = (\epsilon_1, \dots, \epsilon_d) \in G_d \text{ and } \mathbf{x} \in \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_d},$$

which represents the trivial equivalence of singular vector tuples in a concise mathematical way, i.e., all singular vector tuples in the same orbit under the group action by G_d are regarded as one (class). Note that the number of positive singular vector tuples is determined by the matrix tuples $(P_1, \dots, P_d)$ satisfying (P) (cf. (40)), and the group action can be extended to the matrix tuples directly. Let $k \in \{1, \dots, D\}$ be fixed. Then the number of matrix tuples $(P_1, \dots, P_d)$ satisfying (P) is

$$\#(\mathrm{P})(k) := \underbrace{2^{d-1}\cdots 2^{d-1}}_{k} = 2^{k(d-1)}.$$

It is also easy to see that $(\epsilon_1 P_1,\dots,\epsilon_d P_d)$ satisfies (P) as long as $(P_1,\dots,P_d)$ does for any $(\epsilon_1,\dots,\epsilon_d)\in G_d$. Therefore, the number of positive singular vector tuples for this fixed k is (cf. [38])

$$\frac{\#(\mathrm{P})(k)}{\#(G_d)}=\frac{2^{k(d-1)}}{2^{d-1}}=2^{(k-1)(d-1)}.$$

Thus, the total number of positive singular vector tuples [3] is

$$\binom{D}{1}+2^{d-1}\binom{D}{2}+\cdots+2^{(D-1)(d-1)}\binom{D}{D}=\frac{(1+2^{d-1})^D-1}{2^{d-1}}.$$

Note that the above formula depends extremely on the fact that $d>2$, since the number of nonzero singular vector tuples for matrix cases is D, for generic matrices.

With the picture of the group action and the multilinear orthogonal invariance of singular vector tuples (cf. Proposition 6.3), we can simplify the analysis for studying the nondegeneracy of these vast number of nonzero singular vector tuples. Actually, we can apply a multilinear orthogonal transformation as Proposition 6.3 by diagonal orthogonal matrices Q_i's, where Q_i is a diagonal matrix with the (i,i)-th diagonal element the same as that of P_i if $i\in\Lambda_k$ and 1 otherwise. Since P_i's satisfy (P), we have that $\mathcal{A}=(Q_1,\dots,Q_d)\cdot\mathcal{A}$, and more importantly Q_iP_i is nonnegative for all $i\in\{1,\dots,d\}$, such that the resulting singular vector tuples are in the form (40) with nonnegative P_i's. Thus, we only need to study the nondegeneracy of those singular vector tuples in the form (40) with nonnegative P_i's. For easy reference, these singular vector tuples are dubbed as *essential singular vector tuples*. It is easy to see that the number of nonzero singular vector tuples with nonnegative P_i's is

$$\binom{D}{1}+\binom{D}{2}\cdots+\binom{D}{D}=2^D-1.$$

In the following, we show that all these 2^D-1 singular vector tuples are nondegenerate. Firstly, direct computation shows that the Jacobian matrix of the polynomial mapping $S(\mathbf{x})$ induced by the tensor $\mathcal{A}$ in the form of (39) at a singular vector tuple $\mathbf{x}$ with singular value σ has the form

$$\begin{aligned}\nabla_{\mathbf{x}}S(\mathbf{x})&=-\sigma\,\mathrm{diag}\{I,I,\dots,I\}-\sigma\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_d\end{bmatrix}\begin{bmatrix}\mathbf{x}_1^\mathsf{T}\ \mathbf{x}_2^\mathsf{T}\ \dots\ \mathbf{x}_d^\mathsf{T}\end{bmatrix}+\begin{bmatrix}0&\mathcal{A}\tau_{1,2}(\mathbf{x})&\dots&\mathcal{A}\tau_{1,r}(\mathbf{x})\\ \mathcal{A}\tau_{2,1}(\mathbf{x})&0&\dots&\mathcal{A}\tau_{2,r}(\mathbf{x})\\ \vdots&\vdots&\ddots&\vdots\\ \mathcal{A}\tau_{r,1}(\mathbf{x})&\mathcal{A}\tau_{r,2}(\mathbf{x})&\dots&0\end{bmatrix}\\ &=-\sigma\,\mathrm{diag}\{I,I,\dots,I\}-\sigma\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_d\end{bmatrix}\begin{bmatrix}\mathbf{x}_1^\mathsf{T}\ \mathbf{x}_2^\mathsf{T}\ \dots\ \mathbf{x}_d^\mathsf{T}\end{bmatrix}+\begin{bmatrix}0&\sigma P_1P_2^\mathsf{T}&\dots&\sigma P_1P_d^\mathsf{T}\\ \sigma P_2P_1^\mathsf{T}&0&\dots&\sigma P_2P_d^\mathsf{T}\\ \vdots&\vdots&\ddots&\vdots\\ \sigma P_dP_1^\mathsf{T}&\sigma P_dP_2^\mathsf{T}&\dots&0\end{bmatrix},\end{aligned}\quad(41)$$

where the last equality follows from (39), (40) and Property (P). We show $\mathcal{A}\tau_{1,2}(\mathbf{x})=\sigma P_1P_2^\mathsf{T}$ for an illustration. Suppose without loss of generality that $\Lambda_k=\{1,\dots,k\}$. With $\prod_{j=1}^d(P_j)_{ii}=1$ for all $i\in\{1,\dots,k\}$, a direct calculation will give

$$\mathcal{A}\tau_{1,2}(\mathbf{x})=\sum_{i=1}^k\lambda_i\mathbf{e}_i^1\otimes\mathbf{e}_i^2\prod_{j=3}^d\langle\mathbf{e}_i^j,P_j\mathbf{w}\rangle=\sigma\sum_{i=1}^k\mathbf{e}_i^1\otimes\mathbf{e}_i^2\prod_{j=3}^d(P_j)_{ii}=\sigma\sum_{i=1}^k\mathbf{e}_i^1\otimes\mathbf{e}_i^2(P_1)_{ii}(P_2)_{ii},$$

which is the same as $\sigma P_1P_2^\mathsf{T}$.

[3] While finalizing a first version of this work, Professor Bernd Sturmfels kindly pointed out to the authors that the total number of nonzero real singular vector tuples as well as the characterization of the set of nonzero singular vector tuples for an orthogonally decomposable tensor, has also been recently derived in [53] using algebraic geometry tools. It is worth noting that our derivation is more elementary. Moreover, our main concern here is the nondegeneracy, which is not considered in [53].

Without loss of generality (adopting a simultaneous row and column permutation), we can assume that $P_1 = \operatorname{diag}\{I, 0\}$ for the identity matrix $I \in \mathbb{R}^{k\times k}$, i.e., $\Lambda_k = \{1, \dots, k\}$. In this case, it is easy to see that

$$P_1P_i^\mathsf{T} = [I \quad 0] \text{ and } P_iP_j^\mathsf{T} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \text{ for all } i, j > 1,$$

and in which the identity matrices are of the same size $k \times k$ and the zero block matrices are of appropriate sizes. By Schur's complement theory [32], it is easy to see that the nonsingularity of $\nabla_\mathbf{x} S(\mathbf{x})$ in (41) is equivalent to the nonsingularity of the matrix $\overline{A} \in \mathbb{R}^{(\sum_{i=1}^d n_i+1)\times(\sum_{i=1}^d n_i+1)}$ where

$$\overline{A} := \begin{bmatrix} -I & P_1P_2^\mathsf{T} & \dots & P_1P_d^\mathsf{T} & \mathbf{x}_1 \\ P_2P_1^\mathsf{T} & -I & \dots & P_2P_d^\mathsf{T} & \mathbf{x}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ P_dP_1^\mathsf{T} & P_dP_2^\mathsf{T} & \dots & -I & \mathbf{x}_d \\ \mathbf{x}_1^\mathsf{T} & \mathbf{x}_2^\mathsf{T} & \dots & \mathbf{x}_d^\mathsf{T} & 1 \end{bmatrix}.$$

For convenience, it is of advantage to view $\overline{A}$ as an $(d+1)\times(d+1)$ block matrix. It is easy to see that for $i \in \{1, \dots, d\}$, in the i-th block row

$$R = \begin{bmatrix} P_iP_1^\mathsf{T} & P_iP_2^\mathsf{T} & \cdots & -I & \cdots & P_iP_d^\mathsf{T} & \mathbf{x}_i \end{bmatrix} \in \mathbb{R}^{n_i\times(\sum_{i=1}^d n_i+1)},$$

the $(k+1)$-th row to the n_i-th row are zeros except the i-th block is the $(n_i-k)\times(n_i-k)$ minus identity matrix. Pictorially, we have that

$$R = \begin{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} & \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} & \cdots & \begin{bmatrix} -I & 0 \\ 0 & -I \end{bmatrix} & \cdots & \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} \mathbf{w}' \\ 0 \end{bmatrix} \end{bmatrix},$$

where $\mathbf{w}'$ is the nonzero sub-vector of $P_1\mathbf{w}$. We have a similar conclusion on the i-th block column for $i \in \{1, \dots, d\}$. Therefore, we can apply a simultaneous permutation on both the rows and the columns of $\overline{A}$ to obtain a matrix of the form

$$\begin{bmatrix} -I & & & & & \\ & -I & I & \dots & I & \mathbf{w}' \\ & I & -I & \dots & I & \mathbf{w}' \\ & \vdots & \vdots & \ddots & \vdots & \vdots \\ & I & I & \dots & -I & \mathbf{w}' \\ & (\mathbf{w}')^\mathsf{T} & (\mathbf{w}')^\mathsf{T} & \dots & (\mathbf{w}')^\mathsf{T} & 1 \end{bmatrix}.$$

Thus, we can assume without loss of generality that $n_1 = \cdots = n_d$ and the matrices $P_1 = \cdots = P_d = I$ are the same $n_1 \times n_1$ identity matrix at the very beginning. In the following, we will show that the matrix $\overline{A}$ is nonsingular in this case.

Since the $(1,1)$-th block of $\overline{A}$ is $-I$, we can apply first a row and then a column elementary transformations to $\overline{A}$ to get the following new matrix

$$\begin{bmatrix} -I & 0 & & 0\dots & 0 & \mathbf{0} \\ 0 & 0 & 2I & \dots & 2I & 2\mathbf{w} \\ 0 & 2I & 0 & \dots & 2I & 2\mathbf{w} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 2I & 2I & \dots & 0 & 2\mathbf{w} \\ \mathbf{0} & 2\mathbf{w}^\mathsf{T} & 2\mathbf{w}^\mathsf{T} & \dots & 2\mathbf{w}^\mathsf{T} & 2 \end{bmatrix},$$

where we used the fact that $\mathbf{w}^\mathsf{T}\mathbf{w} = 1$. It is then sufficient to show the nonsingularity of the next matrix

$$\begin{bmatrix} 0 & I & \dots & I & \mathbf{w} \\ I & 0 & \dots & I & \mathbf{w} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ I & I & \dots & 0 & \mathbf{w} \\ \mathbf{w}^\mathsf{T} & \mathbf{w}^\mathsf{T} & \dots & \mathbf{w}^\mathsf{T} & 1 \end{bmatrix}.$$

Permuting the first column with the second column and using the $(1,1)$-th block as the pivot and applying the first column to eliminate the rest blocks in the first row, we get that the nonsingularity is equivalent to that of the following matrix C

$$C := \begin{bmatrix} I & I & I & \dots & I & \mathbf{w} \\ I & -I & 0 & \dots & 0 & \mathbf{0} \\ I & 0 & -I & \dots & 0 & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ I & 0 & 0 & \dots & -I & \mathbf{0} \\ \mathbf{w}^\mathsf{T} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & 0 \end{bmatrix}.$$

The matrix C can be naturally divided into a 2×2 block matrix. We first show that the $(1,1)$-th block matrix B is nonsingular. Explicitly,

$$B := \begin{bmatrix} I & I & I & \dots & I \\ I & -I & 0 & \dots & 0 \\ I & 0 & -I & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ I & 0 & 0 & \dots & -I \end{bmatrix}.$$

The nonsingularity of B can be easily shown with Laplace's formula for determinant (cf. [32, Page 28]) by induction starting from the right-down corner block $-I$ and the first block in the last row. Suppose that B is of block size $s \times s$, then it is also easy to see by induction that

$$B^{-1} = \frac{1}{s} B.$$

Finally, the nonsingularity of the matrix C follows from that $B^{-1} = \frac{1}{s}B$, and the fact that $\mathbf{w}^\mathsf{T}\mathbf{w} = 1$ with Schur's complement theory. In conclusion, we see that the singular vector tuple is nondegenerate.

Below, let us summarize these facts for orthogonally decomposable tensors in the following proposition for easy reference.

Proposition 6.5 (Singular Vector Tuples of Orthogonally Decomposable Tensors)**.** *Let $d \geq 3$ and $Q_i = [\mathbf{a}_{i,1}, \dots, \mathbf{a}_{i,n_i}] \in \mathbb{R}^{n_i \times n_i}$ be orthogonal matrices, $i = 1, \dots, d$. Let $\mathcal{A}$ be an orthogonally decomposable tensor with the form $\mathcal{A} = \sum_{i=1}^D \lambda_i \mathbf{a}_{1,i} \otimes \dots \otimes \mathbf{a}_{d,i}$ for some $D \leq \min\{n_1, \dots, n_d\}$, and $\lambda_i > 0$, $i = 1, \dots, D$. Then, there are $\frac{(1+2^{d-1})^D - 1}{2^{d-1}}$ singular vector tuples of $\mathcal{A}$ with positive singular values, and each one of them is nondegenerate and of the form: for some $k \in \{1, \dots, D\}$*

$$(\sigma, \mathbf{x}_1, \dots, \mathbf{x}_d) = \left(\left(\frac{1}{\sum_{i \in \Lambda_k} \lambda_i^{\frac{-2}{d-2}}} \right)^{\frac{d-2}{2}}, Q_1 P_1 \mathbf{w}, \dots, Q_d P_d \mathbf{w} \right),$$

where $\Lambda_k \subseteq \{1, \dots, D\}$ be a subset of cardinality $k \leq D$,

$$\mathbf{w} = \left(\frac{1}{\sum_{i \in \Lambda_k} \lambda_i^{\frac{-2}{d-2}}} \right)^{\frac{1}{2}} \left(\lambda_1^{\frac{-1}{d-2}}, \dots, \lambda_D^{\frac{-1}{d-2}} \right)^\mathsf{T} \in \mathbb{R}^D,$$

and $P_i \in \mathbb{R}^{n_i \times D}$ *is a diagonal matrix with the* (j,j)*-th diagonal element being* 1 *or* -1 *exactly when* $j \in \Lambda_k$ *and zero otherwise such that the product of all the* (j,j)*-th diagonal elements of* P_i*'s with* $i \in \{1,\dots,d\}$ *being one for all* $j \in \Lambda_k$*. Moreover, every nonzero singular vector tuple is nondegenerate.*

The next result is a direct consequence of Theorem 4.4, Remark 3.5 and Proposition 6.5, which shows that HOPM converges globally with R-linear convergence rate for orthogonally decomposable tensors. We emphasize that, in [24], global convergence rate estimates in terms of the tangents of the angles between the iterations and a given point were carefully analyzed for HOPM (or equivalently alternating least squares methods). In particular, under a dominance hypothesis (which involves the information of the generated iterations), a superlinear convergence rate was established in the case of orthogonally decomposable tensors. We refer to [24] and the general version [23] for details. Our next result differs from the result in [24] in the sense that we do not assume any regularity assumptions. Moreover, our approach offers other insights for orthogonally decomposable tensors (such as nondegeneracy of nonzero singular vector tuples) which is of independent interest.

Corollary 6.6 (**Linear convergence of HOPM for orthogonally decomposable Tensors).** *Let* $\{\mathbf{x}^{(k)}\}$ *be generated by Algorithm 3.1 for a given nonzero orthogonally decomposable tensor* $\mathcal{A}$ *with order* $d \geq 3$*. Then* $\mathbf{x}^{(k)} \to \mathbf{x}^*$ *for a nonzero singular vector tuple* $\mathbf{x}^*$*, and* $\{\mathbf{x}^{(k)}\}$ *converges to* $\mathbf{x}^*$ *globally with R-linear convergence rate.*

Acknowledgement

The authors are grateful to the referees and the editor for their constructive comments and helpful suggestions which have contributed to the final presentation of the paper. The authors would also like to thank Dr. Yang Qi (University of Chicago) for pointing out the reference [7].

The first author's work is partially supported by National Science Foundation of China (Grant No. 11771328), Young Elite Scientists Sponsorship Program by Tianjin, and Innovation Research Foundation of Tianjin University (Grant Nos. 2017XZC-0084 and 2017XRG-0015). The second author is partially supported by a Future fellowship from Australian Research Council (FT130100038) and a discovery project from Australian Research Council (DP180100745).

References

[1] A. Ammar, F. Chinesta, A. Falcó, On the convergence of a greedy rank-one update algorithm for a class of linear systems, Arch. Comput. Methods Eng., (2010) 17: 473–486.

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

[3] H. Attouch, J. Bolte, P. Redont, A. Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality, Math. Oper. Res., (2010) 35: 438–457.

[4] A. Beck, First-order methods in optimization, MOS-SIAM Series on Optimization, SIAM, 2017.

[5] J. Bolte, A. Daniilidis, O. Ley, L. Mazet, Characterizations of Łojasiewicz inequalities and applications: subgradient flows, talweg, convexity, Trans. Amer. Math. Soc., (2010) 362: 3319–3363.

[6] W.M. Boothby, An Introduction to Differentiable Manifolds and Riemannian Geometry, Academic Press, New York, 1975.

[7] R. Bott, L.W. Tu, Differential Forms in Algebraic Topology, Springer, 1982.

[8] R. Bro, PARAFAC, tutorial and applications, Chemom. Intell. Lab. Systems, (1997) 38: 149–171.

[9] J.-F. Cardoso, P. Comon, Independent component analysis, a survey of some algebraic methods, in the IEEE International Symposium on Circuits and Systems, Vol. 2, IEEE, New York, 1996, pp. 93–96.

[10] J.-F. Cardoso, A. Souloumiac, Jacobi angles for simultaneous diagonalization, SIAM J. Matrix Anal. Appl., (1996) 17: 161–164.

[11] J.D. Carroll, S. Pruzansky, The CANDECOMP-CANDELINC family of models and methods for multidimensional data analysis, in Research Methods for Multimode Data Analysis, H.G. Law, C.W. Snyder, J.A. Hattie, R.P. McDonald, eds., Praeger, New York, 1984, pp. 372–402.

[12] P. Comon, MA identification using fourth order cumulants, Signal Process., (1992) 26: 381–388.

[13] P. Comon, Independent component analysis, a new concept?, Signal Process., (1994) 36: 287–314.
[14] P. Comon, B. Mourrain, Decomposition of quantics in sums of powers of linear forms, Signal Process., (1996) 53: 93–107.
[15] D. D' Acunto, K. Kurdyka, Explicit bounds for the Łojasiewicz exponent in the gradient inequality for polynomials, Ann. Polon. Math., (2005) 87: 51–61.
[16] L. De Lathauwer, Signal Processing Based on Multilinear Algebra, Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 1997.
[17] L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl., (2000) 21: 1253–1278.
[18] L. De Lathauwer, B. De Moor, J. Vandewalle, On the best rank-1 and rank-$(R_1, R_2, \dots, R_N)$ approximation of higher-order tensors, SIAM J. Matrix Anal. Appl., (2000) 21: 1324–1342.
[19] L. De Lathauwer, P. Comon, B. De Moor, and J. Vandewalle, Higher-order power method - Application in independent component analysis, in Proceedingsof the International Symposium on Nonlinear Theory and ItsApplications(NOLTA'95), LasVegas, NV, (1995), 91-?6.
[20] V. de Silva, L.-H. Lim, Tensor rank and the ill-posedness of the best low-rank approximation problem, SIAM J. Matrix Anal. Appl., (2008) 30: 1084–1127.
[21] M.P. do Carmo, Riemannian Geometry, Springer, 1992.
[22] A. Edelman, T. Arias, S.T. Smith, The geometry of algorithms with orthogonality constraints, SIAM J. Matrix Anal. Appl., (1998) 20: 303–353.
[23] M. Espig, W. Hackbusch, A. Khachatryan, On the convergence of alternating least squares optimization in tensor format representations, arXiv 1506.00062v1, 2015.
[24] M. Espig, A. Khachatryan, Convergence of alternating least squares optimization for rank-one approximation to higher order tensors, arXiv 1503.05431, 2015.
[25] P. Frankel, G. Garrigos, J. Peypouquet, Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates, J. Optim. Theory Appl., (2015) 165: 874–900.
[26] S. Friedland, G, Ottaviani, The number of singular vector tuples and uniqueness of best rank-one approximation of tensors, Found. Comput. Math., (2014) 14: 1209–1242.
[27] I.M. Gelfand, M.M. Kapranov, A.V. Zelevinsky, Discriminants, Resultants and Multidimensional Determinants, Birkhäuser, Boston, 1994.
[28] G.H. Golub, C.F. Van Loan, Matrix Computations, 4th ed., Johns Hopkins University Press, Baltimore, MD, 2013.
[29] L. Grasedyck, D. Kressner, C. Tobler, A literature survey of low-rank tensor approximation techniques, GAMM-Mitt., (2013) 36: 53–78.
[30] W. Hackbusch, Tensor Spaces and Numerical Tensor Calculus, Springer, 2012.
[31] C.J. Hillar, L.-H. Lim, Most tensor problems are NP-hard, J. ACM, (2013) 60: 45, 1–39.
[32] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
[33] E. Kofidis, P. Regalia, On the best rank-1 approximation of higher-order supersymmetric tensors, SIAM J. Matrix Anal. Appl., (2002) 23: 863–884.
[34] T.G. Kolda, Orthogonal tensor decompositions, SIAM J. Matrix Anal. Appl., (2001) 23: 243–255.
[35] T.G. Kolda, B.W. Bader, Tensor decompositions and applications, SIAM Rev., (2009) 51: 455–500.
[36] P.M. Kroonenberg, Three-Mode Principal Component Analysis, DSWO Press, Leiden, The Netherlands, 1983.
[37] J.M. Landsberg, Tensors: Geometry and Applications, AMS, Providence, RI, 2012.
[38] S. Lang, Algebra, revised third edition, Springer, 2004.
[39] G. Li, B.S. Mordukhovich, T.S. Pham, New fractional error bounds for polynomial systems with applications to Hölderian stability in optimization and spectral theory of tensors, Math. Program., (2015) 153: 333–362.
[40] G. Li, B.S. Mordukhovich, T.T.A. Nghia, T.S. Pham, Error bounds for parametric polynomial systems with applications to higher-order stability analysis and convergence rates, Math. Program. (2016). doi:10.1007/s10107-016-1014-6.
[41] G. Li, T.K. Pong, Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods, Found. Comput. Math., 2017, DOI 10.1007/s10208-017-9366-8.
[42] G. Li, L.Q. Qi, G. Yu, Semismoothness of the maximum eigenvalue function of a symmetric tensor and its application, Linear Algebra and its Applications (2013) 438: 813-833.
[43] L.-H. Lim, Singular values and eigenvalues of tensors: a variational approach, in Proceedings of the 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005, pp. 129–132.
[44] S. Łojasiewicz, Une propriété topologique des sous-ensembles analytiques réels, Les Équations aux Dérivées Partielles. Éditions du centre National de la Recherche Scientifique, Paris, 1963, pp. 87–89.
[45] M. J. Mohlenkamp, Musings on multilinear fitting, Linear Algebra Appl., (2013) 438: 834–852.
[46] M. Morse, The calculus of variations in the large, Amer. Math. Soc., 1934.

[47] Y. Nesterov, Introductory Lectures on Convex Optimization - A Basic Course, Kluwer Academic Publishers, 2004.
[48] J. Nie, L. Wang, Semidefinite relaxations for best rank-1 tensor approximations, SIAM J. Matrix Anal. Appl., (2014) 35: 1155–1179.
[49] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Springer, 1970.
[50] L. Qi, Eigenvalues of a real supersymmetric tensor, J. Symbolic Comput., (2005), 40: 1302–1324.
[51] L. Qi, The best rank-one approximation ratio of a tensor space, SIAM J. Matrix Anal. Appl., (2011) 32: 430–442.
[52] E. Robeva, Orthogonal decomposition of symmetric tensors, SIAM J. Matrix Anal. Appl., (2016) 37: 86–102.
[53] E. Robeva, A. Seigal, Singular vectors of orthogonally decomposable tensors, Linear and Multilinear Alge., (2017), http://dx.doi.org/10.1080/03081087.2016.1277508
[54] R.T. Rockafellar, R. Wets, Variational Analysis. Grundlehren der Mathematischen Wissenschaften, Vol. 317. Springer, Berlin, 1998.
[55] R. Schneider, A. Uschmajew, Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality, SIAM J. Optim., (2015) 25: 622–646.
[56] I.R. Shafarevich, Basic Algebraic Geometry, Springer-Verlag, Berlin, 1977.
[57] A. J. Sommese, C.W. Wampler, II, The Numerical Solution of Systems of Polynomials Arising in Engineering and Science, World Scientific, Hackensack, NJ, 2005.
[58] A. Uschmajew, Local convergence of the alternating least squares algorithm for canonical tensor approximation, SIAM J. Matrix Anal. Appl., 2012, 33: 639–652.
[59] A. Uschmajew, A new convergence proof for the high-order power method and generalizations, Pacific J. Optim., (2015) 11: 309–321.
[60] L. Wang, M. Chu, On the global convergence of the alternating least squares method for rank-one approximation to generic tensors, SIAM J. Matrix Anal. Appl., (2014) 23: 1058–1072.
[61] Y. Yang, Y. Feng, X. Huang, J.A.K. Suykens, Rank-1 tensor properties with applications to a class of tensor optimization problems, SIAM J. Optim., (2016) 26: 171–196.
[62] Y. Yang, Y. Feng, J.A.K. Suykens, A rank-one tensor updating algorithm for tensor completion, IEEE Signal Process. Lett., (2015) 22: 1633–1637.
[63] T. Zhang, G.H. Golub, Rank-one approximation to high order tensors, SIAM J. Matrix Anal. Appl., (2001) 23: 534–550.
[64] X. Zhang, C. Ling, L. Qi, The best rank-1 approximation of a symmetric tensor and related spherical optimization problems, SIAM J. Matrix Anal. and Appl., (2012): 33: 806–821.

Department of Mathematics, School of Science, Hangzhou Dianzi University, Hangzhou, 310018, China
E-mail address: shenglonghu@hdu.edu.cn

Department of Applied Mathematics, University of New South Wales, Sydney 2052, Australia.
E-mail address: g.li@unsw.edu.au