arXiv:1504.00538v3 [math.NA] 9 Oct 2017

On the convergence of higher-order orthogonal iteration

Yangyang Xu[a]*

aDepartment of Mathematical Sciences
Rensselaer Polytechnic Institute
110 8th Street, Troy, NY 12180, USA

The higher-order orthogonal iteration (HOOI) has been popularly used for finding a best low-multilinear-rank approximation of a tensor. However, its convergence is still an open question. In this paper, we first analyze a greedy HOOI, which updates each factor matrix by selecting from the best candidates one that is closest to the current iterate. Assuming the existence of a block-nondegenerate cluster point, we establish its global iterate sequence convergence through the so-called Kurdyka-Łojasiewicz (KL) property. In addition, we show that if the starting point is sufficiently close to any block-nondegenerate globally optimal solution, the greedy HOOI produces an iterate sequence convergent to a globally optimal solution. Relating the iterate sequence by the original HOOI to that by the greedy HOOI, we then show that the original HOOI has global convergence on the multilinear subspace sequence and thus positively address the open question.

Keywords: higher-order orthogonal iteration (HOOI), global convergence, Kurdyka-Łojasiewicz (KL) property, greedy algorithm, block coordinate descent

AMS Subject Classification: 9008; 90C26; 90C59;

1. Introduction

It is shown in [1] that any tensor (i.e., multi-dimensional array) can be decomposed into the product of orthogonal matrices and an *all-orthogonal* core tensor. This decomposition generalizes the matrix SVD and is today commonly called higher-order singular value decomposition (HOSVD) or multilinear SVD. In applications, a low-multilinear-rank approximation of a given tensor is commonly used, such as the multilinear subspace learning [2], multilinear principal component analysis [3], tensor decomposition in signal processing [4], just to name a few. Unlike the matrix SVD, *truncated* HOSVD can give a good but not necessarily the best low-multilinear-rank approximation of the given tensor. To obtain a better approximation, existing works (e.g., [5–7]) solve the best rank-$(r_1, \ldots, r_N)$ approximation problem

$$\min_{\mathcal{C}, \mathbf{A}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{C}} \times_1 \mathbf{A}_1 \ldots \times_N \mathbf{A}_N\|_F^2, \text{ s.t. } \mathbf{A}_n \in \text{St}_{I_n \times r_n}, \forall n, \tag{1}$$

*Corresponding author. Email: xuy21@rpi.edu

where $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ is a given tensor, $\times_n$ denotes mode-n tensor-matrix multiplication (see the definition in (3) below), and

$$\mathrm{St}_{I_n \times r_n} = \{\mathbf{A}_n \in \mathbb{R}^{I_n \times r_n} : \mathbf{A}_n^\top \mathbf{A}_n = \mathbf{I}\}$$

is the Stiefel manifold. With $\mathbf{A}$ fixed, the optimal core tensor is given by $\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1^\top \ldots \times_N \mathbf{A}_N^\top$. Absorbing this $\boldsymbol{\mathcal{C}}$ into the objective, one can write (1) equivalently to (see [6, Theorem 3.1] for detailed derivation)

$$\max_{\mathbf{A}} \|\boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1^\top \ldots \times_N \mathbf{A}_N^\top\|_F^2, \text{ s.t. } \mathbf{A}_n \in \mathrm{St}_{I_n \times r_n}, \forall n. \tag{2}$$

One popular method for solving (2) is the higher-order orthogonal iteration (HOOI) (see Algorithm 1). HOOI is commonly used and practically efficient (already coded in the Matlab Tensor Toolbox [8] and Tensorlab [9]). Its iterate sequence convergence has been established for the case of rank-one tensor decomposition [10, 11]. However, for general cases, existing works only show that the objective value of (2) at the generated iterates increasingly converges to some value while the iterate sequence convergence is still an open question (c.f. [1]). In this paper, we address this open question by showing a result on *multilinear subspace convergence*. This result is important because without convergence, running the algorithm to different numbers of iterations may give severely different multilinear subspaces, and that will ultimately affect the results of applications. Our main results are summarized in the following theorem.

THEOREM 1.1 (Main Theorem)   *Let $\{\mathbf{A}^k\}_{k \geq 1}$ be the sequence generated by the HOOI method. We have:*

(i). If $\{\mathbf{A}^k\}_{k \geq 1}$ has a block-nondegenerate (see Definition 1) cluster point $\bar{\mathbf{A}}$, then $\bar{\mathbf{A}}$ is a critical point and also a block-wise maximizer of (2). In addition, $\lim_{k \to \infty} \mathbf{A}^k(\mathbf{A}^k)^\top = \bar{\mathbf{A}}\bar{\mathbf{A}}^\top$, where

$$\mathbf{A}\mathbf{A}^\top = (\mathbf{A}_1\mathbf{A}_1^\top, \ldots, \mathbf{A}_N\mathbf{A}_N^\top).$$

(ii). If the starting point $\mathbf{A}^0$ is sufficiently close to any block-nondegenerate local maximizer of (2), then the entire sequence $\{\mathbf{A}^k(\mathbf{A}^k)^\top\}_{k \geq 1}$ must converge to some point $\bar{\mathbf{A}}\bar{\mathbf{A}}^\top$ and $\bar{\mathbf{A}}$ is a local maximizer of (2).

We make some remarks on the assumption and the convergence results.

Remark 1   The block-nondegeneracy assumption is also necessary because even starting from a critical point $\bar{\mathbf{A}}$, the HOOI method can still deviate from $\bar{\mathbf{A}}$ if it is not block-nondegenerate (see Remark 3), that is, a degenerate critical point is not stable (see [12] for the perturbation analysis). In practice, the block-nondegeneracy is always observed[1], and it is implied by $\liminf_k \left(\sigma_{r_n}(\mathbf{G}_n^k) - \sigma_{r_n+1}(\mathbf{G}_n^k)\right) > 0, \forall n$, where $\mathbf{G}_n^k$ is defined in (6); see Figure 2.

The assumption is similar to the one assumed by the orthogonal iteration method [13, section 7.3.2] for computing r-dimensional dominant invariant subspace of a matrix $\mathbf{X}$. Typically, the convergence of the orthogonal iteration method requires

---

[1]Here, we assume $r_n \leq \prod_{i \neq n} r_i, \forall n$. If otherwise, for some n, $r_n > \prod_{i \neq n} r_i$, we can reduce r_n to $\prod_{i \neq n} r_i$ without changing the approximation in (1) because $\mathbf{G}_n^k \in \mathbb{R}^{I_n \times \prod_{i \neq n} r_i}$.

that there is a positive gap between the r-th and $(r+1)$-th largest eigenvalues of $\mathbf{X}$ in magnitude, because otherwise, the r-dimensional dominant invariant subspace of $\mathbf{X}$ is not unique.

For a block-wise maximizer $\bar{\mathbf{A}}$, its block-nondegeneracy is equivalent to negative definiteness of each block Hessian over the Stiefel manifold $\mathrm{St}_{I_n \times r_n}$. The definition of our block-nondegeneracy is different from the nondegeneracy in [14]. A nondegenerate local maximizer in [14] is one local maximizer that has negative definite Hessian, so the nondegeneracy assumption in [14] is strictly stronger than our block-nondegeneracy assumption.

Remark 2   Since the solution to each subproblem (see (5)) of the HOOI method is not unique and actually still a solution after multiplying any orthogonal matrix to its right, we can only hope to establish convergence of the projection matrix sequence $\{\mathbf{A}^k (\mathbf{A}^k)^\top\}_{k \geq 1}$ instead of $\{\mathbf{A}^k\}_{k \geq 1}$ itself. Hence, the convergence result is on the product manifold of $\mathrm{St}_{I_n \times r_n}$ and similar to that in [14]. However, our assumption is strictly weaker, and thus different tools are used.

1.1.   *Basic concepts of tensor*

Before proceeding with discussion, we first review some basic concepts about tensor that we use in this paper.

The $(i_1, \ldots, i_N)$-th component of an N-way tensor $\boldsymbol{\mathcal{X}}$ is denoted as $x_{i_1 \ldots i_N}$. For $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \in \mathbb{R}^{m_1 \times \cdots \times m_N}$, their inner product is defined in the same way as that for matrices, i.e.,

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{i_1=1}^{m_1} \cdots \sum_{i_N=1}^{m_N} x_{i_1 \ldots i_N} \cdot y_{i_1 \ldots i_N}.$$

The Frobenius norm of $\boldsymbol{\mathcal{X}}$ is defined as $\|\boldsymbol{\mathcal{X}}\|_F = \sqrt{\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle}$. A *fiber* of $\boldsymbol{\mathcal{X}}$ is a vector obtained by fixing all indices of $\boldsymbol{\mathcal{X}}$ except one. The mode-n *matricization* (also called *unfolding*) of $\boldsymbol{\mathcal{X}}$ is denoted as $\mathbf{unfold}_n(\boldsymbol{\mathcal{X}})$, which is a matrix with columns being the mode-n fibers of $\boldsymbol{\mathcal{X}}$ in the lexicographical order. The mode-n product of $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{m_1 \times \cdots \times m_N}$ with $\mathbf{Y} \in \mathbb{R}^{p \times m_n}$ is written as $\boldsymbol{\mathcal{X}} \times_n \mathbf{Y}$ which gives a tensor in $\mathbb{R}^{m_1 \times \cdots \times m_{n-1} \times p \times m_{n+1} \times \cdots \times m_N}$ and is defined component-wisely by

$$(\boldsymbol{\mathcal{X}} \times_n \mathbf{Y})_{i_1 \cdots i_{n-1} j i_{n+1} \cdots i_N} = \sum_{i_n=1}^{m_n} x_{i_1 i_2 \cdots i_N} \cdot y_{j i_n}. \tag{3}$$

If $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_1 \mathbf{A}_1 \ldots \times_N \mathbf{A}_N$, then for any n,

$$\begin{aligned} \mathbf{unfold}_n(\boldsymbol{\mathcal{X}}) &= \mathbf{A}_n \mathbf{unfold}_n(\boldsymbol{\mathcal{C}}) (\mathbf{A}_N \otimes \ldots \otimes \mathbf{A}_{n+1} \otimes \mathbf{A}_{n-1} \otimes \ldots \otimes \mathbf{A}_1)^\top, \\ &= \mathbf{A}_n \mathbf{unfold}_n(\boldsymbol{\mathcal{C}} \times_1 \mathbf{A}_1 \ldots \times_{n-1} \mathbf{A}_{n-1} \times_{n+1} \mathbf{A}_{n+1} \ldots \times_N \mathbf{A}_N), \end{aligned} \tag{4}$$

where "$\otimes$" denotes the Kronecker product.

3

1.2. *Higher-order orthogonal iteration*

The HOOI method updates $\mathbf{A}$ by maximizing the objective of (2) alternatingly with respect to $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N$, one factor matrix at a time while the remaining ones are fixed. Specifically, assuming the iterate to be $\mathbf{A}^k$ at the beginning of the k-th iteration, it performs the following update sequentially from $n = 1$ through N:

$$\mathbf{A}_n^{k+1} \in \underset{\mathbf{A}_n \in \mathrm{St}_{I_n \times r_n}}{\arg\max} \ \|\mathbf{A}_n^\top \mathbf{G}_n^k\|_F^2, \tag{5}$$

where we have used (4), and

$$\mathbf{G}_n^k = \mathbf{unfold}_n(\boldsymbol{\mathcal{X}} \times_{i<n} (\mathbf{A}_i^{k+1})^\top \times_{i>n} (\mathbf{A}_i^k)^\top). \tag{6}$$

Any orthonormal basis of the dominant r_n-dimensional left singular subspace of $\mathbf{G}_n^k$ is a solution of (5). The pseudocode of HOOI is given in Algorithm 1.

Algorithm 1: Higher-order orthogonal iteration (HOOI)

---

1 **Input:** $\boldsymbol{\mathcal{X}}$ and $(r_1, \ldots, r_N)$
2 **Initialization:** choose $(\mathbf{A}_1^0, \ldots, \mathbf{A}_N^0)$ with $\mathbf{A}_n^0 \in \mathrm{St}_{I_n \times r_n}$, $\forall n$
3 **for** $k = 0, \ldots,$ **do**
4   **for** $n = 1, \ldots, N$ **do**
5   Set $\mathbf{A}_n^{k+1}$ to an orthonormal basis of the dominant r_n-dimensional left singular subspace of $\mathbf{G}_n^k$.
6   **if** *Some stopping criteria are met* **then**
7   Output $\mathbf{A} = \mathbf{A}^{k+1}$, $\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1 \ldots \times_N \mathbf{A}_N$ and stop.

It is easy to implement Algorithm 1 by simply setting $\mathbf{A}_n^{k+1}$ to the left r_n leading singular vectors of $\mathbf{G}_n^k$. This implementation is adopted in the Matlab Tensor Toolbox [8] and Tensorlab [9]. However, we did not find any work that gives a convergence result of HOOI, except for our recent paper [15] that establishes subsequence convergence by assuming a strong condition on the entire iterate sequence. The essential difficulty is the non-uniqueness of the solution of (5), and the leading singular vectors are not uniquely determined either.

To tackle this difficulty, we first analyze a greedy method, which always chooses one solution of (5) that is closest to $\mathbf{A}_n^k$ as follows:

$$\mathbf{A}_n^{k+1} \in \underset{\mathbf{A}_n \in \mathcal{H}_n^k}{\arg\min} \|\mathbf{A}_n - \mathbf{A}_n^k\|_F^2, \tag{7}$$

where

$$\mathcal{H}_n^k = \underset{\mathbf{A}_n \in \mathrm{St}_{I_n \times r_n}}{\arg\max} \ \|\mathbf{A}_n^\top \mathbf{G}_n^k\|_F^2. \tag{8}$$

The pseudocode of the greedy implementation is shown in Algorithm 2. The subproblem in (7) can be solved by the method given in Remark 5. Although (7) can in general have multiple solutions, we will show that near any cluster point of the iterate sequence, it must have a unique solution. With the greedy implementation, we are able to establish iterate sequence convergence of the greedy HOOI method

4

(i.e., Algorithm 2), as shown in section 2. Through relating (see (31) and Figure 1) the two iterate sequences generated by the original (i.e., Algorithm 1) and greedy HOOI methods, we then establish the multilinear subspace convergence of the original HOOI method, as shown in section 3.

Algorithm 2: Greedy higher-order orthogonal iteration (Greedy-HOOI)

---

1 **Input:** $\mathcal{X} \in \mathbb{R}^{I_1 \times \dots \times I_N}$ and $(r_1, \dots, r_N)$
2 **Initialization:** choose $(\mathbf{A}_1^0, \dots, \mathbf{A}_N^0)$ with $\mathbf{A}_n^0 \in \mathrm{St}_{I_n \times r_n}$, $\forall n$
3 **for** $k = 0, \dots,$ **do**
4   **for** $n = 1, \dots, N$ **do**
5 $\lfloor$ Set $\mathbf{A}_n^{k+1}$ by (7)
6   **if** *Some stopping criteria are met* **then**
7 $\lfloor$ Output $\mathbf{A} = \mathbf{A}^{k+1}$, $\mathcal{C} = \mathcal{X} \times_1 \mathbf{A}_1 \dots \times_N \mathbf{A}_N$ and stop.

1.3.   *Comparison to other methods*

Besides the HOOI method, several other methods have been developed for solving the low-multilinear-rank tensor approximation problem. One of the earliest methods, called TUCKALS3, was proposed in [5]. TUCKALS3 also sequentially updates $\mathbf{A}_1$ through $\mathbf{A}_N$ and then cycles the process, but different from HOOI, it obtains approximate leading left singular vectors of $\mathbf{G}_n^k$ by carrying out only one step of the so-called Bauer-Rutishauser method [16] starting from $\mathbf{A}_n^k$. This update is equivalent to solving a linearized version of the subproblem (5), and it prevents $\mathbf{A}_n^{k+1}$ being far away from $\mathbf{A}_n^k$. Subsequence convergence of TUCKALS3 was established under the assumption that $(\mathbf{A}_n^k)^\top \mathbf{G}_n^k (\mathbf{G}_n^k)^\top \mathbf{A}_n^k$ is positive definite for all n and k. Although TUCKALS3 has slightly lower per-iteration complexity than HOOI, it does not converge as fast as HOOI as demonstrated in Figure 1.

Recently, some Newton-type methods on manifolds were developed for the low-multilinear-rank tensor approximation problem such as the (quasi)Newton-Grassmann method [7, 17], geometric Newton method [18] and the Riemannian trust region scheme [14]. These methods usually exhibit superlinear convergence. Numerical experiments in [14] demonstrate that for small-size problems, the Riemannian trust region scheme and/or Newton-type methods can take much fewer iterations and also less time than the HOOI method to reach a high-level accuracy based on violation to the first-order optimality condition. However, for medium-size or large-size problems, or if only medium-level accuracy is required, the HOOI method is superior over the Riemannian trust region scheme and also several other Newton-type methods.

Under negative definiteness assumption on the Hessian of a local maximizer, the Newton-type methods are guaranteed to have superlinear or even quadratic local convergence (c.f. [14]). Compared to our block-nondegeneracy assumption, their assumption is strictly stronger because as mentioned in Remark 1, for a local maximizer, its block-nondegeneracy is equivalent to the negative definiteness of each block Hessian. Only with block-nondegeneracy assumption, it is not clear how to show the local convergence of the Newton-type methods.
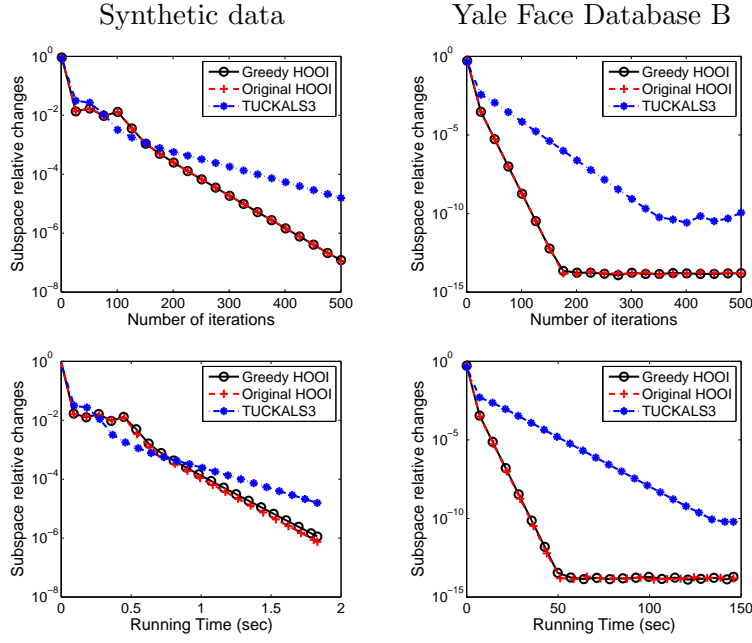
Figure 1.   Comparison of HOOIs and TUCKALS3 [5] on a randomly generated tensor of $50 \times 50 \times 50$ with core size $5 \times 5 \times 5$ and the Yale Face Database B [19, 20] of size $38 \times 64 \times 2958$ with core size $5 \times 5 \times 20$. All three methods start from the same point, which is given by truncated HOSVD. The subspace relative change is calculated by $\frac{\sum_{n=1}^{N} \|\mathbf{A}_n^k (\mathbf{A}_n^k)^\top - \mathbf{A}_n^{k+1} (\mathbf{A}_n^{k+1})^\top \|_F}{\sum_{n=1}^{N} \|\mathbf{A}_n^k (\mathbf{A}_n^k)^\top \|_F}$, and it measures how far the current iterate deviates from satisfying the first-order optimality conditions. The results show that the original HOOI and the greedy HOOI give the same relative change of multilinear subspace at each iteration. This is because they produce the same multilinear subspace. They converge faster than TUCKALS3 on both synthetic data and the face image database.
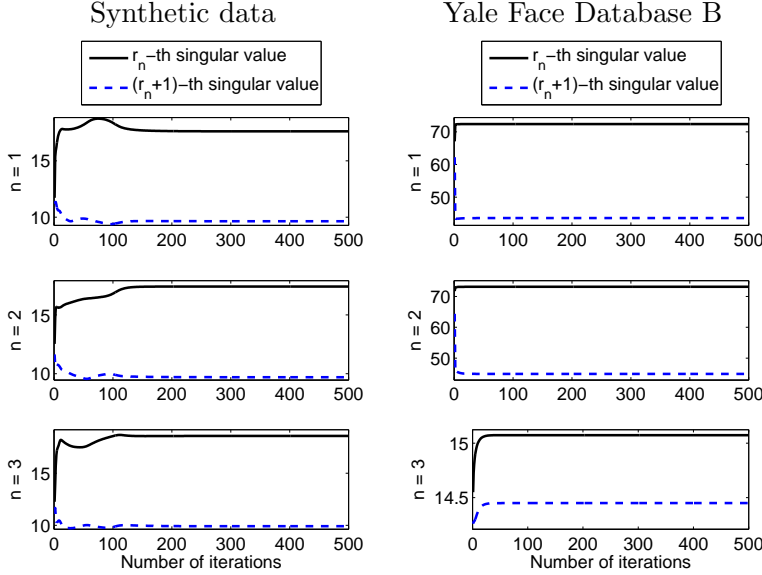

Figure 2.   More observations on the tests in Figure 1: the r_n-th and $(r_n + 1)$-th singular values of $\mathbf{G}_n^k$ generated by the original HOOI on the random dataset and the face image database; the values by the greedy HOOI are the same. Clearly, there are positive gaps between the r_n-th and $(r_n + 1)$-th singular values of $\mathbf{G}_n^k$ for all n in the limit.

6

1.4. *Contributions*

We summarize our contributions as follows.

– We propose a greedy HOOI method, which, for each update, selects from the best candidates one that is closest to the current iterate. With the greedy implementation, we show that any block-nondegenerate cluster point is a critical point and also a block-wise maximizer, and if a block-nondegenerate cluster point exists, then the entire iterate sequence converges to this cluster point.
– Through relating the two iterate sequences by the original and greedy HOOIs, we — for the first time — establish global convergence of the original HOOI on multilinear subspace by assuming the existence of a block-nondegenerate cluster point and local convergence to a locally optimal subspace by assuming sufficient closeness of the starting point to a block-nondegenerate local maximizer.
– As a result, we show that the original HOOI converges to a globally optimal multilinear subspace, if the starting point is sufficiently close to any block-nondegenerate globally optimal solution.

1.5. *Notation and outline*

We use bold capital letters $\mathbf{X}, \mathbf{Y}, \ldots$ to denote matrices, caligraphic letters $\mathcal{S}, \mathcal{U}, \ldots$ for (set-valued) mappings, and bold caligraphic letters $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}, \ldots$ for tensors. $\mathbf{I}$ denotes an identity matrix, whose size is clear from the context. The i-th largest singular value of a matrix $\mathbf{X}$ is denoted by $\sigma_i(\mathbf{X})$. The set of all orthonormal matrices in $\mathbb{R}^{m \times r}$ is denoted as $\text{St}_{m \times r} = \{\mathbf{X} \in \mathbb{R}^{m \times r} : \mathbf{X}^\top \mathbf{X} = \mathbf{I}\}$. Throughout the paper, we focus on real field, but our analysis can be directly extended to complex field.

Definition 1 (block-nondegeneracy)   A feasible solution $\mathbf{A}$ of (2) is block-nondegenerate if $\sigma_{r_n}(\mathbf{G}_n) > \sigma_{r_n+1}(\mathbf{G}_n), \forall n$, where

$$\mathbf{G}_n = \mathbf{unfold}_n(\boldsymbol{\mathcal{X}} \times_{i \neq n} \mathbf{A}_i^\top). \tag{9}$$

Remark 3   In general, we are only able to claim convergence with existence of a block-nondegenerate cluster point. The original HOOI method can deviate from a critical point if it is not block-nondegenerate. To see this, suppose $\bar{\mathbf{A}}$ is a block-wise maximizer and thus a critical point. Assume $\sigma_{r_1}(\bar{\mathbf{G}}_1) = \sigma_{r_1+1}(\bar{\mathbf{G}}_1)$. Let the original HOOI method start from $\bar{\mathbf{A}}$ and update the first factor to $\tilde{\mathbf{A}}_1$. Then $\tilde{\mathbf{A}}_1$ may not span the same subspace as that by $\bar{\mathbf{A}}_1$ because $\bar{\mathbf{G}}_1$ has more than one dominant r_1-dimensional left singular subspaces. Therefore, we cannot guarantee the convergence of the learned multilinear subspace.

The rest of the paper is organized as follows. In section 2, global sequence convergence of the greedy HOOI is established under the assumption of the existence of a block-nondegenerate cluster point. The convergence of the original HOOI is shown in section 3. Finally, section 4 concludes the paper.

7

2.   Convergence analysis

In this section, we assume the existence of one block-nondegenerate cluster point and establish global sequence convergence of Algorithm 2 to a critical point. We first show some properties of the solution to (5). These properties are important to show the block-wise maximality of a cluster point. Then we prove sufficient progress after each iteration of Algorithm 2. Finally we use the so-called Kurdyka-Łojasiewicz (KL) property to establish the global sequence convergence. Note that if $\bar{\mathbf{A}}$ is a critical point of (2), then letting $\bar{\boldsymbol{\mathcal{C}}} = \boldsymbol{\mathcal{X}} \times_1 \bar{\mathbf{A}}_1 \ldots \times_N \bar{\mathbf{A}}_N$, we have $(\bar{\boldsymbol{\mathcal{C}}}, \bar{\mathbf{A}})$ to be a critical point of (1). Hence, our analysis will only focus on (2) and its equivalent form $\max_{\mathbf{A}} F(\mathbf{A})$, where

$$F(\mathbf{A}) = f(\mathbf{A}) - \sum_{n=1}^{N} g_n(\mathbf{A}_n),$$

with $f(\mathbf{A}) = \|\boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1^\top \ldots \times_N \mathbf{A}_N^\top\|_F^2$ and

$$g_n(\mathbf{A}_n) = \begin{cases} 0, & \text{if } \mathbf{A}_n \in \text{St}_{I_n \times r_n}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Since F is a semi-algebraic function, it has the so-called KL property (e.g., see [21]), namely, for any point $\bar{\mathbf{A}}$, in a neighborhood $\mathcal{N}(\bar{\mathbf{A}}, \rho)$, there exists $\phi(s) = cs^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1)$ such that

$$\phi'(|F(\mathbf{A}) - F(\bar{\mathbf{A}})|)\text{dist}(\mathbf{0}, \partial F(\mathbf{A})) \geq 1, \text{ for any } \mathbf{A} \in \mathcal{N}(\bar{\mathbf{A}}, \rho) \text{ and } F(\mathbf{A}) \neq F(\bar{\mathbf{A}}). \tag{10}$$

The KL property was introduced by Łojasiewicz [22] on real analytic functions. Kurdyka extended this property to functions on the o-minimal structure in [23]. Recently, the KL inequality was extended to nonsmooth sub-analytic functions [24]. The works [21, 25] give a lot of concrete examples that own the property.

2.1.   *First-order optimality conditions*

The Lagrangian function of (2) is

$$\mathcal{L}(\mathbf{A}, \boldsymbol{\Lambda}) = \frac{1}{2}\|\boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1^\top \ldots \times_N \mathbf{A}_N^\top\|_F^2 - \frac{1}{2}\sum_{n=1}^{N}\langle \boldsymbol{\Lambda}_n, \mathbf{A}_n^\top \mathbf{A}_n - \mathbf{I}\rangle,$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_N)$ is the Lagrangian multiplier. The KKT conditions or first-order optimality conditions of (2) can be derived from $\nabla\mathcal{L} = \mathbf{0}$, namely,

$$\mathbf{G}_n\mathbf{G}_n^\top \mathbf{A}_n - \mathbf{A}_n\boldsymbol{\Lambda}_n = \mathbf{0}, \forall n, \tag{11a}$$

$$\mathbf{A}_n^\top \mathbf{A}_n - \mathbf{I} = \mathbf{0}, \forall n, \tag{11b}$$

where $\mathbf{G}_n$ is defined in (9). From (11), we have $\boldsymbol{\Lambda}_n = \mathbf{A}_n^\top \mathbf{G}_n\mathbf{G}_n^\top \mathbf{A}_n$. Hence, the condition in (11a) can be written to

$$\mathbf{G}_n\mathbf{G}_n^\top \mathbf{A}_n = \mathbf{A}_n\mathbf{A}_n^\top \mathbf{G}_n\mathbf{G}_n^\top \mathbf{A}_n, \forall n, \tag{11c}$$

The above optimality conditions state that the projection of every block-gradient to the tangent space of the Stiefel manifold is zero. A point $\bar{\mathbf{A}}$ is a critical point of (2) if it satisfies the conditions in (11b) and (11c).

The following result is well known, and we will use it several times in our convergence analysis.

LEMMA 2.1 (von Neumann's Trace Inequality [26])   *For any matrices* $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times p}$, *it holds that*

$$|\langle \mathbf{X}, \mathbf{Y} \rangle| \leq \sum_{i=1}^{\min(m,p)} \sigma_i(\mathbf{X}) \sigma_i(\mathbf{Y}). \tag{12}$$

The inequality (12) *holds with equality if* $\mathbf{X}$ *and* $\mathbf{Y}$ *have the same left and right singular vectors.*

2.2.   *Properties of the solution to* (5)

To show the convergence of Algorithm 2, we analyze the solution of the subproblem (5). The established properties are important to show the block-wise maximality of a cluster point. Problem (5) can be written in the following general form:

$$\min_{\mathbf{Z} \in \mathcal{H}_{\mathbf{Y}}} \|\mathbf{Z} - \mathbf{X}\|_F^2, \tag{13}$$

where $\mathbf{X} \in \mathrm{St}_{m \times r}$ and $\mathbf{Y} \in \mathbb{R}^{m \times p}$ are given, and

$$\mathcal{H}_{\mathbf{Y}} = \underset{\mathbf{Z} \in \mathrm{St}_{m \times r}}{\arg \max} \|\mathbf{Z}^\top \mathbf{Y}\|_F^2. \tag{14}$$

Definition 2 (Quotient set of left leading singular vectors)   Given a matrix $\mathbf{Y} \in \mathbb{R}^{m \times p}$ and positive integer $r \leq \min(m, p)$, define

$\mathcal{B}(\mathbf{Y}, r) = \{\mathbf{U} \in \mathrm{St}_{m \times r} : \mathrm{span}(\mathbf{U}) \text{ is a dominant } r\text{-dimensional left singular subspace of } \mathbf{Y}\}.$

For any $\mathbf{U}_1, \mathbf{U}_2 \in \mathcal{B}(\mathbf{Y}, r)$, if $\mathrm{span}(\mathbf{U}_1) = \mathrm{span}(\mathbf{U}_2)$, i.e., they span the same subspace, we say they are equivalent. By this equivalence relation, we partition $\mathcal{B}(\mathbf{Y}, r)$ to a set of equivalence classes and form a quotient set denoted as $\mathcal{U}(\mathbf{Y}, r)$.

Remark 4   Throughout the paper, we regard $\mathcal{U}(\mathbf{Y}, r)$ as the finite set of orthonormal matrices, and each of its elements is a representative of the bases that span the same subspace. If $\sigma_r(\mathbf{Y}) > \sigma_{r+1}(\mathbf{Y})$, then $\mathbf{Y}$ has a unique dominant r-dimensional left singular subspace, and $\mathcal{U}(\mathbf{Y}, r)$ is a singleton. Otherwise if $\sigma_r(\mathbf{Y}) = \sigma_{r+1}(\mathbf{Y})$, then $\mathbf{Y}$ has multiple dominant r-dimensional left singular subspaces, and $\mathcal{U}(\mathbf{Y}, r)$ has more than one element.

PROPOSITION 2.2   *The problem* (13) *has a unique solution if and only if the following two conditions hold:*

(1) If $\mathbf{U}_* \in \arg \max_{\mathbf{U} \in \mathcal{U}(\mathbf{Y}, r)} \|\mathbf{U}^\top \mathbf{X}\|_*$, *then* $\mathbf{U}_*^\top \mathbf{X}$ *is nonsingular;*
(2) For any $\mathbf{U} \in \mathcal{U}(\mathbf{Y}, r)$, *if* $\mathbf{U} \neq \mathbf{U}_*$, *then* $\|\mathbf{U}^\top \mathbf{X}\|_* < \|\mathbf{U}_*^\top \mathbf{X}\|_*$;

where $\|\cdot\|_*$ denotes matrix nuclear norm, defined as the sum of all singular values of a matrix.

Proof. The "only if" part is easy to see, so we only prove the "if" part. Assume $\tilde{\mathbf{Z}}$ and $\hat{\mathbf{Z}}$ are both solutions of (13). Note that $\mathcal{H}_{\mathbf{Y}}$ in (14) is exactly the set $\mathcal{B}(\mathbf{Y}, r)$. Hence, $\tilde{\mathbf{Z}} = \mathbf{U}_{\tilde{z}}\mathbf{W}_{\tilde{z}}$ and $\hat{\mathbf{Z}} = \mathbf{U}_{\hat{z}}\mathbf{W}_{\hat{z}}$ for $\mathbf{U}_{\tilde{z}}, \mathbf{U}_{\hat{z}} \in \mathcal{U}(\mathbf{Y}, r)$ and some $\mathbf{W}_{\tilde{z}}, \mathbf{W}_{\hat{z}} \in \mathrm{St}_{r \times r}$. Note

$$\|\tilde{\mathbf{Z}} - \mathbf{X}\|_F^2 = 2r - 2\langle\tilde{\mathbf{Z}}, \mathbf{X}\rangle = 2r - 2\langle\mathbf{W}_{\tilde{z}}, \mathbf{U}_{\tilde{z}}^\top\mathbf{X}\rangle.$$

Then by Lemma 2.1 and the optimality of $\tilde{\mathbf{Z}}$ on solving (13), we have

$$\langle\mathbf{W}_{\tilde{z}}, \mathbf{U}_{\tilde{z}}^\top\mathbf{X}\rangle = \sum_{i=1}^r \sigma_i(\mathbf{U}_{\tilde{z}}^\top\mathbf{X}) = \|\mathbf{U}_{\tilde{z}}^\top\mathbf{X}\|_* = \max_{\mathbf{U}\in\mathcal{U}(\mathbf{Y},r)} \|\mathbf{U}^\top\mathbf{X}\|_*. \qquad (15)$$

Hence, from items 1 and 2, it follows that $\mathbf{U}_{\tilde{z}} = \mathbf{U}_*$, and similarly $\mathbf{U}_{\hat{z}} = \mathbf{U}_*$.

Let $\mathbf{U}_*^\top\mathbf{X} = \bar{\mathbf{U}}\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^\top$ be the full SVD of $\mathbf{U}_*^\top\mathbf{X}$ and $\mathbf{V}_{\tilde{z}} = \mathbf{W}_{\tilde{z}}^\top\bar{\mathbf{U}}$, so $\mathbf{W}_{\tilde{z}} = \bar{\mathbf{U}}\mathbf{V}_{\tilde{z}}^\top$. Then from (15), it holds that

$$\sum_{i=1}^r \sigma_i(\mathbf{U}_*^\top\mathbf{X}) = \langle\mathbf{W}_{\tilde{z}}, \mathbf{U}_*^\top\mathbf{X}\rangle = \langle\mathbf{V}_{\tilde{z}}^\top, \bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^\top\rangle = \langle\mathbf{V}_{\tilde{z}}^\top\bar{\mathbf{V}}, \bar{\mathbf{\Sigma}}\rangle = \sum_{i=1}^r \sigma_i(\mathbf{U}_*^\top\mathbf{X})(\mathbf{V}_{\tilde{z}}^\top\bar{\mathbf{V}})_{ii}.$$

Note that $\sigma_i(\mathbf{U}^\top\mathbf{X}) > 0$ and $(\mathbf{V}_{\tilde{z}}^\top\bar{\mathbf{V}})_{ii} \leq 1$. The equality $\sum_{i=1}^r \sigma_i(\mathbf{U}_*^\top\mathbf{X}) = \sum_{i=1}^r \sigma_i(\mathbf{U}_*^\top\mathbf{X})(\mathbf{V}_{\tilde{z}}^\top\bar{\mathbf{V}})_{ii}$ holds only if $(\mathbf{V}_{\tilde{z}}^\top\bar{\mathbf{V}})_{ii} = 1$. Since $\mathbf{V}_{\tilde{z}}^\top\bar{\mathbf{V}}$ is orthogonal, we must have $\mathbf{V}_{\tilde{z}}^\top\bar{\mathbf{V}} = \mathbf{I}$. Hence, $\mathbf{V}_{\tilde{z}} = \bar{\mathbf{V}}$ and $\mathbf{W}_{\tilde{z}} = \bar{\mathbf{U}}\bar{\mathbf{V}}^\top$. For the same reason, $\mathbf{W}_{\hat{z}} = \bar{\mathbf{U}}\bar{\mathbf{V}}^\top$. Therefore, $\tilde{\mathbf{Z}} = \hat{\mathbf{Z}}$, and the solution of (13) is unique. $\square$

Definition 3 (Unique-solution mapping)   Let

$$\mathcal{S}(\mathbf{Y}, r) = \{\mathbf{X} \in \mathrm{St}_{m\times r} : \mathbf{X} \text{ satisfies the two conditions in Proposition 2.2}\}.$$

For any $\mathbf{X} \in \mathcal{S}(\mathbf{Y}, r)$, define $T_{\mathbf{Y},r}(\mathbf{X})$ as the unique solution of (13).

Remark 5   The proof of Proposition 2.2 provides a way for finding a solution of (13). Find $\mathbf{U}_* \in \arg\max_{\mathbf{U}\in\mathcal{U}(\mathbf{Y},r)} \|\mathbf{U}^\top\mathbf{X}\|_*$ and get full SVD of $\mathbf{U}_*^\top\mathbf{X} = \bar{\mathbf{U}}\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^\top$. Then $\mathbf{Z}_* = \mathbf{U}_*\bar{\mathbf{U}}\bar{\mathbf{V}}^\top$ is a solution of (13).

Using Proposition 2.2, one can easily show the following two corollaries.

COROLLARY 2.3   *If $\mathbf{X}$ is sufficiently close to one $\mathbf{U}$ in $\mathcal{B}(\mathbf{Y}, r)$, then the solution of (13) is unique.*

COROLLARY 2.4   *If $\mathbf{X} \in \mathcal{B}(\mathbf{Y}, r)$, then $T_{\mathbf{Y},r}(\mathbf{X}) = \mathbf{X}$, i.e., $\mathbf{X}$ is a fixed point.*

Furthermore, we can show the continuity of $T_{\mathbf{Y},r}$.

THEOREM 2.5   *The mapping $T_{\mathbf{Y},r}$ is continuous on $\mathcal{S}(\mathbf{Y}, r)$.*

Proof. For convenience of the description, in this proof, we simply write $\mathcal{U}(\mathbf{Y}, r), \mathcal{S}(\mathbf{Y}, r)$ and $T_{\mathbf{Y},r}$ to $\mathcal{U}, \mathcal{S}$ and T, respectively.

For any $\mathbf{X} \in \mathcal{S}$, let $\mathbf{Z} = T(\mathbf{X})$. If T is not continuous at $\mathbf{X}$, then there exists $\epsilon > 0$ and a sequence $\{\mathbf{X}^k\}_{k\geq 1}$ in $\mathcal{S}$ such that $\|\mathbf{X} - \mathbf{X}^k\|_F \leq \frac{1}{k}$ and $\|\mathbf{Z} - \mathbf{Z}^k\|_F \geq \epsilon$,

where $\mathbf{Z}^k = T(\mathbf{X}^k)$. By the definition of $\mathcal{S}$, we know that there is $\mathbf{U} \in \mathcal{U}$ such that $\|\mathbf{U}^\top \mathbf{X}\|_* > \|\tilde{\mathbf{U}}^\top \mathbf{X}\|_*$ for any $\tilde{\mathbf{U}} \in \mathcal{U}\backslash\{\mathbf{U}\}$. Similarly, there is a sequence $\{\mathbf{U}^k\}_{k \geq 1}$ in $\mathcal{U}$ such that for each k, $\|(\mathbf{U}^k)^\top \mathbf{X}^k\|_* > \|\tilde{\mathbf{U}}^\top \mathbf{X}^k\|_*$ for any $\tilde{\mathbf{U}} \in \mathcal{U}\backslash\{\mathbf{U}^k\}$.

Let $\delta = \|\mathbf{U}^\top \mathbf{X}\|_* - \max_{\tilde{\mathbf{U}} \in \mathcal{U}\backslash\{\mathbf{U}\}} \|\tilde{\mathbf{U}}^\top \mathbf{X}\|_* > 0$. There is a sufficiently large integer k_0 such that for all $k \geq k_0$, it holds $\|\mathbf{U}^\top \mathbf{X}^k\|_* \geq \|\mathbf{U}^\top \mathbf{X}\|_* - \frac{\delta}{4}$ and $\|(\mathbf{U}^k)^\top \mathbf{X}^k\|_* \leq \|(\mathbf{U}^k)^\top \mathbf{X}\|_* + \frac{\delta}{4}$. Note $\|\mathbf{U}^\top \mathbf{X}^k\|_* \leq \|(\mathbf{U}^k)^\top \mathbf{X}^k\|_*$. Hence, $\|\mathbf{U}^\top \mathbf{X}\|_* - \frac{\delta}{4} \leq \|(\mathbf{U}^k)^\top \mathbf{X}\|_* + \frac{\delta}{4}$, i.e., $\|\mathbf{U}^\top \mathbf{X}\|_* \leq \|(\mathbf{U}^k)^\top \mathbf{X}\|_* + \frac{\delta}{2}$. Therefore, by the definition of δ, it must hold that $\mathbf{U}^k = \mathbf{U}$, $\forall k \geq k_0$.

Hence, we can write $\mathbf{Z} = \mathbf{U}\mathbf{W}_z$ and $\mathbf{Z}^k = \mathbf{U}\mathbf{W}_{z^k}$ for all $k \geq k_0$, where $\mathbf{W}_z, \mathbf{W}_{z^k} \in \mathrm{St}_{r \times r}$. Note $\mathbf{U}^\top \mathbf{X}^k \to \mathbf{U}^\top \mathbf{X}$ as $k \to \infty$. Then from the proof of Proposition 2.2, we have $\mathbf{W}_{z^k} \to \mathbf{W}_z$ and thus $\mathbf{Z}^k \to \mathbf{Z}$ as $k \to \infty$. This contradicts to $\|\mathbf{Z} - \mathbf{Z}^k\|_F \geq \epsilon$. Therefore, T is continuous at $\mathbf{X}$. Since $\mathbf{X}$ is an arbitrary point in $\mathcal{S}$, this completes the proof. $\qquad\square$

One can also show the following result.

THEOREM 2.6   *Assume $\sigma_r(\mathbf{Y}) > \sigma_{r+1}(\mathbf{Y})$ and $\mathbf{Y}^k \to \mathbf{Y}$ as $k \to \infty$. If $\mathbf{X} \in \mathcal{S}(\mathbf{Y}, r)$, then there is a sufficiently large integer k_0 such that $\mathbf{X} \in \mathcal{S}(\mathbf{Y}^k, r)$ for all $k \geq k_0$, and*

$$\lim_{k \to \infty} T_{\mathbf{Y}^k, r}(\mathbf{X}) = T_{\mathbf{Y}, r}(\mathbf{X}). \tag{16}$$

Proof. By the assumption $\sigma_r(\mathbf{Y}) > \sigma_{r+1}(\mathbf{Y})$, $\mathcal{U}(\mathbf{Y}, r)$ is a singleton. Let $\mathbf{U} \in \mathcal{U}(\mathbf{Y}, r)$. Then from $\mathbf{X} \in \mathcal{S}(\mathbf{Y}, r)$, it follows that $\mathbf{X}^\top \mathbf{U}$ is nonsingular. Since $\mathbf{Y}^k \to \mathbf{Y}$ as $k \to \infty$, there exists an integer k_0, such that $\sigma_r(\mathbf{Y}^k) > \sigma_{r+1}(\mathbf{Y}^k)$, i.e., $\mathcal{U}(\mathbf{Y}^k, p)$ is a singleton for all $k \geq k_0$. Let $\mathbf{U}^k \in \mathcal{U}(\mathbf{Y}^k, r)$, $\forall k$. We can choose the representative satisfying $\mathbf{U}^k \to \mathbf{U}$, since $\mathbf{Y}^k \to \mathbf{Y}$. Therefore, taking another larger k_0 if necessary, we have that $\mathbf{X}^\top \mathbf{U}^k$ is nonsingular and thus $\mathbf{X} \in \mathcal{S}(\mathbf{Y}^k, r)$ for all $k \geq k_0$. Finally, using Remark 5 and $\mathbf{U}^k \to \mathbf{U}$, we have (16) and complete the proof. $\qquad\square$

We also need the following result, which will be used to show the criticality and block-wise maximality of a cluster point of the sequence given by Algorithm 2.

LEMMA 2.7   *For any feasible solution $\bar{\mathbf{A}}$, if $T_{\bar{\mathbf{G}}_n, r_n}(\bar{\mathbf{A}}_n) = \bar{\mathbf{A}}_n$, $\forall n$, then $\bar{\mathbf{A}}$ is a critical point and also a block-wise maximizer of (2), where*

$$\bar{\mathbf{G}}_n = \mathbf{unfold}_n(\boldsymbol{\mathcal{X}} \times_{i \neq n} \bar{\mathbf{A}}_i). \tag{17}$$

Proof. Note that $T_{\bar{\mathbf{G}}_n, r_n}(\bar{\mathbf{A}}_n) = \bar{\mathbf{A}}_n$, $\forall n$ implies that $\bar{\mathbf{A}}_n$ is a basis of the dominant r_n-dimensional left singular subspace of $\bar{\mathbf{G}}_n$. Hence, $\bar{\mathbf{A}}_n \bar{\mathbf{A}}_n^\top \bar{\mathbf{G}}_n \bar{\mathbf{G}}_n^\top \bar{\mathbf{A}}_n = \bar{\mathbf{G}}_n \bar{\mathbf{G}}_n^\top \bar{\mathbf{A}}_n$, $\forall n$. Therefore, $\bar{\mathbf{A}}$ is a critical point.

In addition, $T_{\bar{\mathbf{G}}_n, r_n}(\bar{\mathbf{A}}_n) = \bar{\mathbf{A}}_n$, $\forall n$ implies that $\bar{\mathbf{A}}_n$ is a solution to $\max_{\mathbf{A}_n} \|\mathbf{A}_n^\top \bar{\mathbf{G}}_n\|_F^2$ over $\mathrm{St}_{I_n \times r_n}$ for all n. Hence, $\bar{\mathbf{A}}$ is a block-wise maximizer. This completes the proof. $\qquad\square$

2.3.   *Bounding iterate distance by objective progress*

As shown below, for the problem (13), if there is a positive gap between $\sigma_r(\mathbf{Y})$ and $\sigma_{r+1}(\mathbf{Y})$, the distance between $\mathbf{X}$ and $\mathbf{Z}$ can be bounded by the objective difference.

THEOREM 2.8   *Given* $\mathbf{X} \in \mathrm{St}_{m \times r}$ *and* $\mathbf{Y} \in \mathbb{R}^{m \times p}$, *any solution* $\mathbf{Z}$ *of* (13) *satisfies*

$$\frac{\sigma_r^2(\mathbf{Y}) - \sigma_{r+1}^2(\mathbf{Y})}{2} \|\mathbf{Z} - \mathbf{X}\|_F^2 \leq \|\mathbf{Z}^\top \mathbf{Y}\|_F^2 - \|\mathbf{X}^\top \mathbf{Y}\|_F^2. \tag{18}$$

Proof. Note $\mathbf{Z} = \mathbf{U}\mathbf{W}_z$ for some $\mathbf{U} \in \mathcal{U}(\mathbf{Y}, r)$ and $\mathbf{W}_z \in \mathrm{St}_{r \times r}$. Let $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top + \mathbf{U}_\perp \mathbf{\Sigma}_\perp \mathbf{V}_\perp^\top$ be the full SVD of $\mathbf{Y}$. Also, let $\mathbf{W} = \mathbf{U}^\top \mathbf{X}$ and $\mathbf{W}_\perp = \mathbf{U}_\perp^\top \mathbf{X}$. Then $\mathbf{X} = \mathbf{U}\mathbf{W} + \mathbf{U}_\perp \mathbf{W}_\perp$ and $\mathbf{W}^\top \mathbf{W} + \mathbf{W}_\perp^\top \mathbf{W}_\perp = \mathbf{I}$ from $\mathbf{X}^\top \mathbf{X} = \mathbf{I}$.

As in the proof of Proposition 2.2, we have

$$\|\mathbf{Z}^\top \mathbf{Y}\|_F^2 = \sum_{i=1}^r \sigma_i^2(\mathbf{Y}) \tag{19}$$

and

$$\|\mathbf{Z} - \mathbf{X}\|_F^2 = 2r - 2\langle \mathbf{Z}, \mathbf{X} \rangle = 2r - 2\langle \mathbf{W}_z, \mathbf{W} \rangle = 2r - 2\sum_{i=1}^r \sigma_i(\mathbf{W}), \tag{20}$$

where the last equality is from Lemma 2.1 and the optimality of $\mathbf{Z}$ for (13). Also, note that

$$\|\mathbf{X}^\top \mathbf{Y}\|_F^2 = \|\mathbf{W}^\top \mathbf{\Sigma}\|_F^2 + \|\mathbf{W}_\perp^\top \mathbf{\Sigma}_\perp\|_F^2. \tag{21}$$

Assume $\mathbf{W}_\perp = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^\top$ to be the full SVD of $\mathbf{W}_\perp$. Then

$$\mathbf{W}^\top \mathbf{W} = \mathbf{I} - \mathbf{W}_\perp^\top \mathbf{W}_\perp = \tilde{\mathbf{V}}(\mathbf{I} - \tilde{\mathbf{\Sigma}}^\top \tilde{\mathbf{\Sigma}})\tilde{\mathbf{V}}^\top.$$

Let $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \ldots \geq \tilde{\sigma}_r$ be the first r largest singular values of $\mathbf{W}_\perp$. Then $\sigma_i(\mathbf{W}) = \sqrt{1 - \tilde{\sigma}_{r-i+1}^2}$, $\forall i$, and using Lemma 2.1 again, we have

$$\|\mathbf{W}^\top \mathbf{\Sigma}\|_F^2 = \langle \mathbf{W}\mathbf{W}^\top, \mathbf{\Sigma}^2 \rangle \leq \sum_{i=1}^r (1 - \tilde{\sigma}_i^2)\sigma_{r-i+1}^2(\mathbf{Y}), \tag{22}$$

and

$$\|\mathbf{W}_\perp^\top \mathbf{\Sigma}_\perp\|_F^2 = \langle \mathbf{W}_\perp \mathbf{W}_\perp^\top, \mathbf{\Sigma}_\perp \mathbf{\Sigma}_\perp^\top \rangle \leq \sum_{i=1}^r \tilde{\sigma}_i^2 \sigma_{r+i}^2(\mathbf{Y}). \tag{23}$$

Hence, from (19) and (21) through (23), we have

$$\begin{aligned}
\|\mathbf{Z}^\top \mathbf{Y}\|_F^2 - \|\mathbf{X}^\top \mathbf{Y}\|_F^2 &= \sum_{i=1}^r \sigma_i^2(\mathbf{Y}) - \|\mathbf{W}^\top \mathbf{\Sigma}\|_F^2 - \|\mathbf{W}_\perp^\top \mathbf{\Sigma}_\perp\|_F^2 \\
&\geq \sum_{i=1}^r \tilde{\sigma}_i^2 \left(\sigma_{r-i+1}^2(\mathbf{Y}) - \sigma_{r+i}^2(\mathbf{Y}) \right) \\
&\geq \sum_{i=1}^r \tilde{\sigma}_i^2 \left(\sigma_r^2(\mathbf{Y}) - \sigma_{r+1}^2(\mathbf{Y}) \right),
\end{aligned} \tag{24}$$

where the last inequality is from $\sigma_r^2(\mathbf{Y}) - \sigma_{r+1}^2(\mathbf{Y}) \leq \sigma_{r-i+1}^2(\mathbf{Y}) - \sigma_{r+i}^2(\mathbf{Y})$, $\forall i$. Using the fact $1 - \sqrt{1-x} \leq x$, $\forall x \in [0, 1]$, we have

$$2r - 2\sum_{i=1}^r \sigma_i(\mathbf{W}) = 2r - 2\sum_{i=1}^r \sqrt{1 - \tilde{\sigma}_i^2} \leq 2\sum_{i=1}^r \tilde{\sigma}_i^2,$$

and thus from (20), it follows that

$$\|\mathbf{Z} - \mathbf{X}\|_F^2 \leq 2\sum_{i=1}^r \tilde{\sigma}_i^2.$$

Plugging the above inequality into (24), we have the desired result. $\square$

Using Theorem 2.8, we show the following result of sufficient progress.

LEMMA 2.9 (Sufficient progress)   *Let $\{\mathbf{A}^k\}_{k\geq 1}$ be the sequence generated from Algorithm 2. Assume it has a block-nondegenerate cluster point $\bar{\mathbf{A}}$. Then there is a constant α such that if $\mathbf{A}^k$ is sufficiently close to $\bar{\mathbf{A}}$, we have*

$$\alpha\|\mathbf{A}^{k+1} - \mathbf{A}^k\|_F^2 \leq F(\mathbf{A}^{k+1}) - F(\mathbf{A}^k).$$

Proof. It is easy to see that there exists a small positive number δ such that if $\|\mathbf{A} - \bar{\mathbf{A}}\|_F \leq \delta$, then

$$\sigma_{r_n}(\mathbf{G}_n) - \sigma_{r_n+1}(\mathbf{G}_n) \geq \frac{1}{2}(\sigma_{r_n}(\bar{\mathbf{G}}_n) - \sigma_{r_n+1}(\bar{\mathbf{G}}_n)) \triangleq \alpha_n > 0, \forall n,$$

where the strict inequality is from the block-nondegeneracy of $\bar{\mathbf{A}}$. Assume $\mathbf{A}^k$ is sufficiently close to $\bar{\mathbf{A}}$ such that

$$\sum_{n=1}^N \sqrt{\frac{2(F(\bar{\mathbf{A}}) - F(\mathbf{A}^k))}{\alpha_n}} + \|\mathbf{A}^k - \bar{\mathbf{A}}\|_F \leq \delta.$$

From Theorem 2.8, it follows that

$$\frac{\alpha_1}{2}\|\mathbf{A}_1^{k+1} - \mathbf{A}_1^k\|_F^2 \leq \|(\mathbf{A}_1^{k+1})^\top \mathbf{G}_1^k\|_F^2 - \|(\mathbf{A}_1^k)^\top \mathbf{G}_1^k\|_F^2 \leq F(\bar{\mathbf{A}}) - F(\mathbf{A}^k),$$

where $\mathbf{G}_1^k$ is defined in (6), and we have used (26). Hence, $\|\mathbf{A}_1^{k+1} - \mathbf{A}_1^k\|_F \leq \sqrt{\frac{2(F(\bar{\mathbf{A}})-F(\mathbf{A}^k))}{\alpha_1}}$ and

$$\|(\mathbf{A}_1^{k+1}, \mathbf{A}_{>1}^k) - \bar{\mathbf{A}}\|_F \leq \|\mathbf{A}_1^{k+1} - \mathbf{A}_1^k\|_F + \|\mathbf{A}^k - \bar{\mathbf{A}}\|_F \leq \delta.$$

Repeating the above arguments, in general, we have for all n that

$$\|\mathbf{A}_n^{k+1} - \mathbf{A}_n^k\|_F \leq \sqrt{\frac{2(F(\bar{\mathbf{A}}) - F(\mathbf{A}^k))}{\alpha_n}},$$

13

and

$$\|(\mathbf{A}_{\leq n}^{k+1}, \mathbf{A}_{>n}^{k}) - \bar{\mathbf{A}}\|_F \leq \sum_{i=1}^{n} \|\mathbf{A}_i^{k+1} - \mathbf{A}_i^k\|_F + \|\mathbf{A}^k - \bar{\mathbf{A}}\|_F \leq \delta.$$

Therefore, every intermediate point $(\mathbf{A}_{\leq n}^{k+1}, \mathbf{A}_{>n}^{k})$ is in $\mathcal{N}(\bar{\mathbf{A}}, \delta) \triangleq \{\mathbf{A} : \|\mathbf{A} - \bar{\mathbf{A}}\|_F \leq \delta\}$, and thus for all n,

$$\frac{\alpha_n}{2} \|\mathbf{A}_n^{k+1} - \mathbf{A}_n^k\|_F^2 \leq \|(\mathbf{A}_n^{k+1})^\top \mathbf{G}_n^k\|_F^2 - \|(\mathbf{A}_n^k)^\top \mathbf{G}_n^k\|_F^2.$$

Let $\alpha = \min_n \frac{\alpha_n}{2} > 0$. Summing the above inequality from $n = 1$ to N gives the desired result. □

2.4. *Global sequence convergence result*

Using Lemma 2.9 and the KL property of F, we show the global sequence convergence of Algorithm 2 to a critical point. Our proof follows two steps. In the first step, we show criticality of any cluster point; in the second step, we apply the claim made in [27]: for problem $\max_{\mathbf{x}} \Phi(\mathbf{x})$, if the sequence $\{\mathbf{x}^k\}_{k \geq 1}$ generated by a certain algorithm satisfies the following two properties

(1) *Sufficient progress*: there is a constant $\rho_1 > 0$, such that $\rho_1 \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \leq \Phi(\mathbf{x}^{k+1}) - \Phi(\mathbf{x}^k)$, $\forall k$;
(2) *Subgradient lower bound*: there is a constant $\rho_2 > 0$ such that for any k, for some $\mathbf{g}^{k+1} \in \partial \Phi(\mathbf{x}^{k+1})$, it holds $\|\mathbf{g}^{k+1}\| \leq \rho_2 \|\mathbf{x}^{k+1} - \mathbf{x}^k\|$,

then the KL property of Φ implies that $\{\mathbf{x}^k\}$ is a Cauchy sequence.

THEOREM 2.10 (Global sequence convergence)   *Let* $\{\mathbf{A}^k\}_{k \geq 1}$ *be the sequence generated from Algorithm 2. If* $\{\mathbf{A}^k\}_{k \geq 1}$ *has a block-nondegenerate cluster point* $\bar{\mathbf{A}}$, *then* $\bar{\mathbf{A}}$ *is a critical point and block-wise maximizer of* (2), *and*

$$\lim_{k \to \infty} \mathbf{A}^k = \bar{\mathbf{A}}. \tag{25}$$

Proof. We first show the criticality and block-wise maximality of $\bar{\mathbf{A}}$. Suppose that $\bar{\mathbf{A}}$ is one block-nondegenerate Since $\bar{\mathbf{A}}$ is a cluster point, there is a subsequence $\{\mathbf{A}^k\}_{k \in \mathcal{K}}$ convergent to $\bar{\mathbf{A}}$. From the update rule in (5), it is easy to see

$$\|(\mathbf{A}_n^{k+1})^\top \mathbf{G}_n^k\|_F^2 \leq \|\bar{\mathbf{A}}_n^\top \bar{\mathbf{G}}_n\|_F^2, \ \forall k, \ \forall n. \tag{26}$$

We claim that $\bar{\mathbf{A}}_1$ is a solution of $\max_{\mathbf{A}_1 \in \mathrm{St}_{I_1 \times r_1}} \|\mathbf{A}_1^\top \bar{\mathbf{G}}_1\|_F^2$. Otherwise, $\|\bar{\mathbf{A}}_1^\top \bar{\mathbf{G}}_1\|_F^2 < \sum_{i=1}^{r_1} \sigma_i^2(\bar{\mathbf{G}}_1)$. Note

$$\lim_{\mathcal{K} \ni k \to \infty} \|(\mathbf{A}_1^{k+1})^\top \mathbf{G}_1^k\|_F^2 = \lim_{\mathcal{K} \ni k \to \infty} \sum_{i=1}^{r_1} \sigma_i^2(\mathbf{G}_1^k) = \sum_{i=1}^{r_1} \sigma_i^2(\bar{\mathbf{G}}_1),$$

which contradicts to (26). Hence, $T_{\bar{\mathbf{G}}_1, r_1}(\bar{\mathbf{A}}_1) = \bar{\mathbf{A}}_1$.

Note that $\mathbf{G}_1^k \to \bar{\mathbf{G}}_1$ as $\mathcal{K} \ni k \to \infty$ and $\mathbf{A}_1^k \in \mathcal{S}(\mathbf{G}_1^k, r_1)$ as $k \in \mathcal{K}$ is sufficiently large. From the block-nondegeneracy of $\bar{\mathbf{A}}$ and Theorems 2.5 and 2.6, we have

$$\lim_{\mathcal{K} \ni k \to \infty} \mathbf{A}_1^{k+1} = \lim_{\mathcal{K} \ni k \to \infty} T_{\mathbf{G}_1^k, r_1}(\mathbf{A}_1^k) = T_{\bar{\mathbf{G}}_1, r_1}(\bar{\mathbf{A}}_1) = \bar{\mathbf{A}}_1. \tag{27}$$

Hence, taking a sufficiently large $k \in \mathcal{K}$, we can make $\|\mathbf{A}_1^{k+1} - \mathbf{A}_1^k\|_F$ sufficiently small, and thus we can repeat the above arguments for $n = 2, \ldots, N$ to conclude

$$\bar{\mathbf{A}}_n \in \underset{\mathbf{A}_n \in \mathrm{St}_{I_n \times r_n}}{\arg \max} \|\mathbf{A}_n^\top \bar{\mathbf{G}}_n\|_F^2, \, \forall n.$$

Therefore, from the definition of $T_{\bar{\mathbf{G}}_n, r_n}$, it holds that $T_{\bar{\mathbf{G}}_n, r_n}(\bar{\mathbf{A}}_n) = \bar{\mathbf{A}}_n, \, \forall n$, and $\bar{\mathbf{A}}$ is a critical point and a block-wise maximizer of (2) from Lemma 2.7.

Note that there is a constant L such that

$$\|\nabla_{\mathbf{A}_n} f(\tilde{\mathbf{A}}) - \nabla_{\mathbf{A}_n} f(\hat{\mathbf{A}})\|_F \leq L\|\tilde{\mathbf{A}} - \hat{\mathbf{A}}\|_F, \, \forall \tilde{\mathbf{A}}, \hat{\mathbf{A}} \in \mathcal{O}, \, \forall n, \tag{28}$$

where

$$\mathcal{O} = \{\mathbf{A} : \mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_N), \, \mathbf{A}_n \in \mathrm{St}_{I_n \times r_n}, \, \forall n\}.$$

For any $n = 1, \ldots, N$, from the optimality of $\mathbf{A}_n^k$ on problem $\max_{\mathbf{A}_n} F(\mathbf{A}_{i<n}^k, \mathbf{A}_n, \mathbf{A}_{i>n}^{k-1})$, it holds that

$$\begin{aligned}
&\mathbf{0} \in \partial_{\mathbf{A}_n} F(\mathbf{A}_{i \leq n}^k, \mathbf{A}_{i>n}^{k-1}) \\
\Leftrightarrow &\mathbf{0} \in \nabla_{\mathbf{A}_n} f(\mathbf{A}_{i \leq n}^k, \mathbf{A}_{i>n}^{k-1}) + \partial g_n(\mathbf{A}_n^k) \\
\Leftrightarrow &\nabla_{\mathbf{A}_n} f(\mathbf{A}^k) - \nabla_{\mathbf{A}_n} f(\mathbf{A}_{i \leq n}^k, \mathbf{A}_{i>n}^{k-1}) \in \nabla_{\mathbf{A}_n} f(\mathbf{A}^k) + \partial g_n(\mathbf{A}_n^k).
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathrm{dist}(\mathbf{0}, \partial F(\mathbf{A}^k)) &\leq \sum_{n=1}^N \|\nabla_{\mathbf{A}_n} f(\mathbf{A}^k) - \nabla_{\mathbf{A}_n} f(\mathbf{A}_{i \leq n}^k, \mathbf{A}_{i>n}^{k-1})\|_F \\
&\overset{(28)}{\leq} NL\|\mathbf{A}^k - \mathbf{A}^{k-1}\|_F,
\end{aligned} \tag{29}$$

which together with Lemma 2.9 indicates that $\{\mathbf{A}^k\}_{k \geq 1}$ owns the two properties stated in the beginning of this subsection. In addition, F has the KL-property in (10), and thus $\{\mathbf{A}^k\}_{k \geq 1}$ is a Cauchy sequence and converges. Since $\bar{\mathbf{A}}$ is a cluster point, then $\mathbf{A}^k \to \bar{\mathbf{A}}$ as $k \to \infty$. This completes the proof. □

Remark 6 The result in (27) is a key step to have the criticality and block-wise maximality. In general, without the block-nondegeneracy assumption, it may not hold.

As long as the starting point is sufficiently close to any block-nondegenerate local maximizer, Algorithm 2 will yield an iterate sequence convergent to a local maximizer as summarized below.

THEOREM 2.11 (Convergence to local minimizer) *Assume Algorithm 2 starts from any point $\mathbf{A}^0$ that is sufficiently close to one block-nondegenerate local maximizer $\mathbf{A}^*$ of $F(\mathbf{A})$. Then the sequence $\{\mathbf{A}^k\}_{k\geq 1}$ converges to a local maximizer.*

Proof. First, note that if some $\mathbf{A}^{k_0}$ is sufficiently close to $\mathbf{A}^*$ and $F(\mathbf{A}^{k_0}) = F(\mathbf{A}^*)$, then $\mathbf{A}^{k_0}$ must also be a local maximizer and block-nondegenerate. In this case, $\mathbf{A}^k = \mathbf{A}^{k_0}$, $\forall k \geq k_0$. Hence, without loss of generality, we can assume $F(\mathbf{A}^k) < F(\mathbf{A}^*)$, $\forall k$. Secondly, note that in the proof of Theorem 2.10, we only use $F(\mathbf{A}^k) < F(\bar{\mathbf{A}})$ and the sufficient closeness of $\mathbf{A}^0$ to $\bar{\mathbf{A}}$ to show $\{\mathbf{A}^k\}_{k\geq 1}$ to be a Cauchy sequence. Therefore, repeating the same arguments, we can show that if $\mathbf{A}^0$ is sufficiently close to $\mathbf{A}^*$, then $\{\mathbf{A}^k\}_{k\geq 1}$ is a Cauchy sequence and thus converges to a block-nondegenerate point $\bar{\mathbf{A}}$ near $\mathbf{A}^*$. From Theorem 2.10, it follows that $\bar{\mathbf{A}}$ is a critical point. We claim $F(\bar{\mathbf{A}}) = F(\mathbf{A}^*)$, i.e., $\bar{\mathbf{A}}$ is a local maximizer. If otherwise $F(\bar{\mathbf{A}}) < F(\mathbf{A}^*)$, then by the KL inequality, it holds that $\phi'(F(\mathbf{A}^*) - F(\bar{\mathbf{A}}))\mathrm{dist}(\mathbf{0}, \partial F(\bar{\mathbf{A}})) \geq 1$, which contradicts to $\mathbf{0} \in \partial F(\bar{\mathbf{A}})$. Hence, $F(\bar{\mathbf{A}}) = F(\mathbf{A}^*)$. This completes the proof. $\qquad\square$

From Theorem 2.11, we can easily get the following local convergence to a globally optimal solution.

THEOREM 2.12 (Global optimality) *Assume Algorithm 2 starts from any point $\mathbf{A}^0$ that is sufficiently close to one block-nondegenerate globally optimal solution $\mathbf{A}^*$ of (2). Then the sequence $\{\mathbf{A}^k\}_{k\geq 1}$ converges to a globally optimal solution.*

3. Proof of the main theorem

In this section, we analyze the convergence of the original HOOI method by relating its iterate sequence to that of the greedy HOOI method. Because any solution to each subproblem of the original HOOI method is still a solution after arbitrary rotation, we do not hope to establish convergence on the iterate sequence $\{\mathbf{A}^k\}_{k\geq 1}$ itself. Instead, we show the convergence of the projection matrix sequence $\{\mathbf{A}^k(\mathbf{A}^{\bar{k}})^\top\}_{k\geq 1}$.

First note that

$$\|\boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1^\top \ldots \times_N \mathbf{A}_N^\top\|_F^2 = \left\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \times_1 (\mathbf{A}_1\mathbf{A}_1^\top) \ldots \times (\mathbf{A}_N\mathbf{A}_N^\top) \right\rangle. \qquad (30)$$

We also need the following two lemmas.

LEMMA 3.1 *If $\bar{\mathbf{A}}\bar{\mathbf{A}}^\top = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top$ and $\tilde{\mathbf{A}}$ is a critical point of (2), then $\bar{\mathbf{A}}$ is also a critical point.*

Proof. Since $\tilde{\mathbf{A}}$ is a critical point of (2), it holds that $\tilde{\mathbf{G}}_n\tilde{\mathbf{G}}_n^\top\tilde{\mathbf{A}}_n = \tilde{\mathbf{A}}_n\tilde{\mathbf{A}}_n^\top\tilde{\mathbf{G}}_n\tilde{\mathbf{G}}_n^\top\tilde{\mathbf{A}}_n$ and $\tilde{\mathbf{A}}_n^\top\tilde{\mathbf{A}}_n = \mathbf{I}$ for all n. Note that $\bar{\mathbf{A}}\bar{\mathbf{A}}^\top = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top$ implies $\bar{\mathbf{G}}_n\bar{\mathbf{G}}_n^\top = \tilde{\mathbf{G}}_n\tilde{\mathbf{G}}_n^\top$. Hence, for any n,

$$\bar{\mathbf{G}}_n\bar{\mathbf{G}}_n^\top\bar{\mathbf{A}}_n\bar{\mathbf{A}}_n^\top = \tilde{\mathbf{G}}_n\tilde{\mathbf{G}}_n^\top\tilde{\mathbf{A}}_n\tilde{\mathbf{A}}_n^\top = \tilde{\mathbf{A}}_n\tilde{\mathbf{A}}_n^\top\tilde{\mathbf{G}}_n\tilde{\mathbf{G}}_n^\top\tilde{\mathbf{A}}_n\tilde{\mathbf{A}}_n^\top = \bar{\mathbf{A}}_n\bar{\mathbf{A}}_n^\top\bar{\mathbf{G}}_n\bar{\mathbf{G}}_n^\top\bar{\mathbf{A}}_n\bar{\mathbf{A}}_n^\top.$$

Multiplying $\bar{\mathbf{A}}_n$ to both sides and noting $\bar{\mathbf{A}}_n^\top\bar{\mathbf{A}}_n = \mathbf{I}$ gives

$$\bar{\mathbf{G}}_n\bar{\mathbf{G}}_n^\top\bar{\mathbf{A}}_n = \bar{\mathbf{A}}_n\bar{\mathbf{A}}_n^\top\bar{\mathbf{G}}_n\bar{\mathbf{G}}_n^\top\bar{\mathbf{A}}_n, \ \forall n,$$

and thus $\bar{\mathbf{A}}$ is a critical point. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

LEMMA 3.2   *Let* $\{\mathbf{A}^k\}_{k\geq 1}$ *be the sequence generated by the original HOOI method and assume it has a block-nondegenerate cluster point* $\bar{\mathbf{A}}$*. If for some* k_0*,* $F(\mathbf{A}^{k_0}) = F(\bar{\mathbf{A}})$*, then there is an integer* $K \geq k_0$ *such that* $\mathbf{A}^k(\mathbf{A}^k)^\top = \mathbf{A}^K(\mathbf{A}^K)^\top$*,* $\forall k \geq K$*.*

Proof. Because $F(\mathbf{A}^k)$ is nondecreasing and upper bounded, we have $\lim_{k\to\infty} F(\mathbf{A}^k) = F(\bar{\mathbf{A}})$ and $F(\mathbf{A}^k) \leq F(\bar{\mathbf{A}})$, so if $F(\mathbf{A}^{k_0}) = F(\bar{\mathbf{A}})$, then $F(\mathbf{A}^k) = F(\bar{\mathbf{A}})$, $\forall k \geq k_0$.

Since $\bar{\mathbf{A}}$ is a cluster point, there must be an integer $K \geq k_0$ such that $\mathbf{A}^K$ is sufficiently close to $\bar{\mathbf{A}}$ and $\mathbf{A}^K$ is block-nondegenerate. Hence, $\mathbf{G}_1^K$ has a unique dominant r_1-dimensional left singular subspace. Note

$$\max_{\mathbf{A}_1 \in \mathrm{St}_{I_1 \times r_1}} \|\mathbf{A}_1^\top \mathbf{G}_1^K\|_F^2 = \sum_{i=1}^{r_1} \sigma_i^2(\mathbf{G}_1^K) = F(\bar{\mathbf{A}}) = \|(\mathbf{A}_1^K)^\top \mathbf{G}_1^K\|_F^2.$$

Therefore, $\mathbf{A}_1^K$ and $\mathbf{A}_1^{K+1}$ both span the dominant r_1-dimensional left singular subspace of $\mathbf{G}_1^K$, and thus $\mathbf{A}_1^{K+1}(\mathbf{A}_1^{K+1})^\top = \mathbf{A}_1^K(\mathbf{A}_1^K)^\top$. Using (30), we can repeat the arguments to have $\mathbf{A}_n^{K+1}(\mathbf{A}_n^{K+1})^\top = \mathbf{A}_n^K(\mathbf{A}_n^K)^\top$, $\forall n$, i.e., $\mathbf{A}^{K+1} = \mathbf{A}^K$. Now starting from $\mathbf{A}^{K+1}$ and repeating the arguments, we have the desired result. $\qquad\square$

By Lemma 3.2, without loss of generality, we assume $F(\mathbf{A}^k) < F(\bar{\mathbf{A}})$, $\forall k$ in the remaining analysis. With Lemmas 3.1 and 3.2, we are now ready to prove the main theorem.

Proof of Theorem 1.1. **Part (i):** Since $\bar{\mathbf{A}}$ is a cluster point of $\{\mathbf{A}^k\}$, there is a subsequence $\{\mathbf{A}^k\}_{k\in\mathcal{K}}$ convergent to $\bar{\mathbf{A}}$, and there is $k_0 \in \mathcal{K}$ such that $\mathbf{A}^{k_0}$ is sufficiently close to $\bar{\mathbf{A}}$. Without loss of generality, we assume that $\mathbf{A}^0$ is sufficiently close to $\bar{\mathbf{A}}$ because otherwise we can set $\mathbf{A}^{k_0}$ as a new starting point and the convergence of $\{\mathbf{A}^k\}_{k\geq 1}$ is equivalent to that of $\{\mathbf{A}^k\}_{k\geq k_0}$. Let $\{\tilde{\mathbf{A}}^k\}$ be the sequence generated by the greedy HOOI method starting from $\tilde{\mathbf{A}}^0 = \mathbf{A}^0$. We go to show that if $\mathbf{A}^0$ is sufficiently close to $\bar{\mathbf{A}}$, then

$$\mathbf{A}^k(\mathbf{A}^k)^\top = \tilde{\mathbf{A}}^k(\tilde{\mathbf{A}}^k)^\top, \ \forall k \geq 1. \tag{31}$$

Repeating the same arguments in the proof of Lemma 2.9, we have that if $\mathbf{A}^0$ is sufficiently close to $\bar{\mathbf{A}}$, then $\tilde{\mathbf{A}}^1$ is also sufficiently close to $\bar{\mathbf{A}}$. Note that when $\mathbf{A}^0$ is sufficiently close to $\bar{\mathbf{A}}$, it is block-nondegenerate and $\sigma_{r_1}(\mathbf{G}_1^0) > \sigma_{r_1+1}(\mathbf{G}_1^0)$. Hence, $\mathbf{A}_1^1$ and $\tilde{\mathbf{A}}_1^1$ both span the dominant r_1-dimensional left singular subspace of $\mathbf{G}_1^0$ and thus $\mathbf{A}_1^1(\mathbf{A}_1^1)^\top = \tilde{\mathbf{A}}_1^1(\tilde{\mathbf{A}}_1^1)^\top$. Since both $\tilde{\mathbf{A}}^0$ and $\tilde{\mathbf{A}}^1$ are sufficiently close to $\bar{\mathbf{A}}$, we have $\sigma_{r_2}(\tilde{\mathbf{G}}_2^0) > \sigma_{r_2+1}(\tilde{\mathbf{G}}_2^0)$. Note $\mathbf{G}_2^0(\mathbf{G}_2^0)^\top = \tilde{\mathbf{G}}_2^0(\tilde{\mathbf{G}}_2^0)^\top$. Hence, $\mathbf{A}_2^1$ and $\tilde{\mathbf{A}}_2^1$ both span the dominant r_2-dimensional left singular subspace of $\mathbf{G}_2^0(\mathbf{G}_2^0)^\top$ and thus $\mathbf{A}_2^1(\mathbf{A}_2^1)^\top = \tilde{\mathbf{A}}_2^1(\tilde{\mathbf{A}}_2^1)^\top$. Repeating the above arguments, we have $\mathbf{A}_n^1(\mathbf{A}_n^1)^\top = \tilde{\mathbf{A}}_n^1(\tilde{\mathbf{A}}_n^1)^\top$, $\forall n$, i.e., $\mathbf{A}^1(\mathbf{A}^1)^\top = \tilde{\mathbf{A}}^1(\tilde{\mathbf{A}}^1)^\top$.

Assume that for some integer $K \geq 1$, it holds $\mathbf{A}^k(\mathbf{A}^k)^\top = \tilde{\mathbf{A}}^k(\tilde{\mathbf{A}}^k)^\top$ and $\tilde{\mathbf{A}}^k \in \mathcal{N}(\bar{\mathbf{A}}, \rho)$ for all $k \leq K$, where ρ is sufficiently small and plays the same role as that in the proof of Theorem 2.10. From (30), it follows that $F(\tilde{\mathbf{A}}^k) = F(\mathbf{A}^k) < F(\bar{\mathbf{A}})$, $\forall k \geq K$. Through the same arguments as those in the proof of Theorem 2.10, we have $\tilde{\mathbf{A}}^{K+1} \in \mathcal{N}(\bar{\mathbf{A}}, \rho)$, and thus $\mathbf{A}^{K+1}(\mathbf{A}^{K+1})^\top = \tilde{\mathbf{A}}^{K+1}(\tilde{\mathbf{A}}^{K+1})^\top$ by the

above arguments that show $\mathbf{A}^1(\mathbf{A}^1)^\top = \tilde{\mathbf{A}}^1(\tilde{\mathbf{A}}^1)^\top$. By induction, we have the result in (31).

Taking another subsequence if necessary, we can assume $\{\tilde{\mathbf{A}}^k\}_{k \in \mathcal{K}}$ converging to $\tilde{\mathbf{A}}$ and thus $\bar{\mathbf{A}}\bar{\mathbf{A}}^\top = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top$ by (31). Note that the block-nondegeneracy of $\tilde{\mathbf{A}}$ is equivalent to that of $\bar{\mathbf{A}}$. Hence, $\tilde{\mathbf{A}}$ is block-nondegenerate and is a critical point and a block-wise maximizer, and $\tilde{\mathbf{A}}^k$ converges to $\tilde{\mathbf{A}}$ by Theorem 2.10. Therefore, $\mathbf{A}^k(\mathbf{A}^k)^\top$ converges to $\bar{\mathbf{A}}\bar{\mathbf{A}}^\top$. From Lemma 3.1, we have that $\bar{\mathbf{A}}$ is a critical point of (2), and from (30), $\bar{\mathbf{A}}$ is a block-wise maximizer. This completes the proof of part (i).

Part (ii): Let $\{\tilde{\mathbf{A}}^k\}_{k \geq 1}$ be the sequence generated by the greedy HOOI method starting from $\tilde{\mathbf{A}}^0 = \mathbf{A}^0$. From Theorem 2.11, it follows that $\tilde{\mathbf{A}}^k$ converges to a local maximizer $\tilde{\mathbf{A}}$ of (2). In addition, by similar arguments as those in the proof of part (i), we can show that (31) still holds. Hence, $\mathbf{A}^k(\mathbf{A}^k)^\top$ converges to $\tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top$, and this completes the proof. $\qquad\square$

4.   Conclusions

We proposed a greedy HOOI method and established its iterate sequence convergence by assuming existence of a block-nondegenerate cluster point. Through relating the iterates by the original HOOI to those by the greedy HOOI, we have shown the global convergence of the HOOI method on multilinear subspace sequence. In addition, if the starting point is sufficiently close to any block-nondegenerate locally optimal point, we showed that the original HOOI could guarantee convergence to a locally optimal multilinear subspace.

References

[1] De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. SIAM Journal on Matrix Analysis and Applications. 2000;21(4):1253–1278.

[2] Lu H, Plataniotis KN, Venetsanopoulos AN. A survey of multilinear subspace learning for tensor data. Pattern Recognition. 2011;44(7):1540–1551.

[3] Lu H, Plataniotis KN, Venetsanopoulos AN. MPCA: Multilinear principal component analysis of tensor objects. Neural Networks, IEEE Transactions on. 2008;19(1):18–39.

[4] Cichocki A, Mandic D, Phan A, Caiafa C, Zhou G, Zhao Q, De Lathauwer L. Tensor decompositions for signal processing applications: from two-way to multiway component analysis. IEEE Signal Processing Magazine. 2015;32(2):145–163.

[5] Kroonenberg PM, De Leeuw J. Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika. 1980;45(1):69–97.

[6] De Lathauwer L, De Moor B, Vandewalle J. On the best rank-1 and rank-$(r_1, r_2, \ldots, r_n)$ approximation of higher-order tensors. SIAM Journal on Matrix Analysis and Applications. 2000;21(4):1324–1342.

[7] Eldén L, Savas B. A newton-grassmann method for computing the best multilinear rank-(r_1, r_2, r_3) approximation of a tensor. SIAM Journal on Matrix Analysis and applications. 2009;31(2):248–271.

[8] Bader BW, Kolda TG, et al. Matlab tensor toolbox version 2.5. 2012 January.

[9] Sorber L, Barel MV, Lathauwer LD. Tensorlab v2.0. Available online. 2014 January; Available from: http://www.tensorlab.net.

[10] Wang L, Chu MT. On the global convergence of the alternating least squares method for rank-one approximation to generic tensors. SIAM Journal on Matrix Analysis and Applications. 2014;35(3):1058–1072.

[11] Uschmajew A. A new convergence proof for the higher-order power method and generalizations. Pacific Journal Of Optimization. 2015;11(2):309–321.

[12] Eldén L, Savas B. Perturbation theory and optimality conditions for the best multilinear rank approximation of a tensor. SIAM Journal on Matrix Analysis and Applications. 2011;32(4):1422–1450.

[13] Golub GH, Van Loan CF. Matrix computations. 3rd ed. Johns Hopkins Studies in the Mathematical Sciences; Baltimore, MD: Johns Hopkins University Press; 1996.

[14] Ishteva M, Absil PA, Van Huffel S, De Lathauwer L. Best low multilinear rank approximation of higher-order tensors based on the Riemannian trust-region scheme. SIAM Journal on Matrix Analysis and Applications. 2011;32(1):115–135.

[15] Xu Y. Fast algorithms for higher-order singular value decomposition from incomplete data. Journal of Computational Mathematics. 2017;35(4):395–420.

[16] Rutishauser H. Computational aspects of FL bauer's simultaneous iteration method. Numerische Mathematik. 1969;13(1):4–13.

[17] Savas B, Lim LH. Quasi-newton methods on grassmannians and multilinear approximations of tensors. SIAM Journal on Scientific Computing. 2010;32(6):3352–3393.

[18] Ishteva M, De Lathauwer L, Absil PA, Van Huffel S. Differential-geometric newton method for the best rank-(r 1, r 2, r 3) approximation of tensors. Numerical Algorithms. 2009;51(2):179–194.

[19] Georghiades AS, Belhumeur PN, Kriegman D. From few to many: Illumination cone models for face recognition under variable lighting and pose. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2001;23(6):643–660.

[20] Lee KC, Ho J, Kriegman D. Acquiring linear subspaces for face recognition under variable lighting. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2005;27(5):684–698.

[21] Xu Y, Yin W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM Journal on Imaging Sciences. 2013;6(3):1758–1789.

[22] Łojasiewicz S. Sur la géométrie semi-et sous-analytique. Ann Inst Fourier (Grenoble). 1993;43(5):1575–1595.

[23] Kurdyka K. On gradients of functions definable in o-minimal structures. In: Annales de l'Institut Fourier; Vol. 48. Chartres: L'Institut, 1950-; 1998. p. 769–784.

[24] Bolte J, Daniilidis A, Lewis A. The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization. 2007;17(4):1205–1223.

[25] Attouch H, Bolte J, Redont P, Soubeyran A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality. Mathematics of Operations Research. 2010;35(2):438–457.

[26] Mirsky L. A trace inequality of John von Neumann. Monatshefte für Mathematik. 1975;79(4):303–306.

[27] Bolte J, Sabach S, Teboulle M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming. 2014;146(1-2):459–494.