ORTHOGONAL RANK-TWO TENSOR APPROXIMATION: A MODIFIED HIGH-ORDER POWER METHOD AND ITS CONVERGENCE ANALYSIS

DRAFT AS OF February 25, 2013

LIQI WANG* AND MOODY T. CHU†

Abstract. With the notable exceptions that tensors of order 2, that is, matrices always have best approximations of arbitrary low ranks and that tensors of any order always have the best rank-one approximation, it is known that high-order tensors can fail to have best low rank approximations. When the condition of orthogonality is imposed, even at the most general case that only one pair of components in the decomposed rank-one tensors are required to be perpendicular, the situation is changed completely — orthogonal low rank approximations always exist. The purpose of this paper is to discuss the best rank-two approximation subject to orthogonality. The conventional high-order power method is modified to address the orthogonality and a rigorous analysis of convergence is provided in this paper.

Key words. tensor decomposition, low rank approximation, alternating least squares, high-order power method, convergence analysis, singular value decomposition

AMS subject classifications. 15A69, 41A50, 68W25, 65J10

1. Introduction.

A real-valued tensor of order k can be represented by a k-way array

$$T = [\tau_{i_1,\ldots,i_k}] \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_k}$$

with elements $\tau_{i_1,\ldots,i_k}$ accessed via k indices. A tensor of the form

$$\mathbf{u}^{(1)} \otimes \ldots \otimes \mathbf{u}^{(k)} := \left[u_{i_1}^{(1)} \ldots u_{i_k}^{(k)} \right],$$

where elements are the products of entries from vectors $\mathbf{u}^{(j)} \in \mathbb{R}^{I_j}$, $j = 1, \ldots, k$, is said to be of rank one. These vectors are referred to as components of the rank-1 tensor [17]. By a tensor decomposition we refer to the rewriting of the given tensor T as the sum of some rank-1 tensors. Depending on the desirable structure, the summation itself may assume different formats resulting in different types of decompositions. The two most prevalent types are the Tucker decomposition [7, 8, 31]

$$T = \sum_{j_1,j_2,\ldots,j_k} g_{j_1,j_2,\ldots,j_k} \mathbf{u}_{j_1}^{(1)} \otimes \ldots \otimes \mathbf{u}_{j_k}^{(k)} \tag{1.1}$$

and the CP decomposition [11, 12, 13, 15]

$$T = \sum_{j} \mathbf{u}_{j}^{(1)} \otimes \ldots \otimes \mathbf{u}_{j}^{(k)}. \tag{1.2}$$

Any of these decomposition finds a wide range of applications. So that we may stay focused on our main point, we only mention without giving details that decomposing a tensor, especially when subject to the constraint of maintaining minimum number of terms in the summation, amounts to finding real-valued solutions for a high-degree polynomial system of equations, which certainly is not a trivial task. Tensor decomposition, including the determination of the tensor rank, has been a major ongoing research topic with numerous results available in the literature.

Instead of demanding a complete decomposition, an alternative approach that has also attracted considerable interest in the field is by ways of low rank approximation. The basic setup consists of two parts. Firstly, the number of terms in the summation is to be fixed a priori, whose choice is a difficult problem in itself [5, 19, 20].

*Department of Mathematics, Dalian University of Technology, (lizzy@mail.dlut.edu.cn)

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205. (chu@math.ncsu.edu) This research was supported in part by the National Science Foundation under grant DMS-1014666.

Secondly, the difference between the given T and the summation is to be minimized in the sense of Frobenius norm. In contrast to the algebraic problem of finding an exact decomposition which is NP-hard, the analytic problem of finding an approximation seems more computationally feasible. Nonetheless, it must be pointed out that, except for the case of order-2 tensors which are simply matrices or the case of rank-1 approximation for general tensors, the best low rank approximation for high-order tensors may not exist at all [9, 19, 21]. An example of such a failure for rank-2 approximation to an order-k tensor, $k > 2$, is given in the next section.

On the other hand, the best low rank approximation (under the Frobenius norm) for matrices is given by the truncated singular value decomposition according to the Eckart-Young theorem. The notion of the best rank-1 approximation for tensors has also been discussed extensively in the literature [2, 8, 9, 16, 29, 36]. In a recent paper, we furnish a rigorous proof for the global convergence of the alternating least squares (ALS) method for the rank-1 approximation [33].

The purpose of this paper is to consider the rank-2 approximation under the added condition of orthogonality. Our main contribution is on proposing a modification of the conventional high-order power method to address the orthogonality while maintaining a monotone ascending property. We then justify the modification by offering a convergence analysis.

This paper is organized as follows. The problem of rank-2 approximation is described in Section 2 where we also introduce the notions of adjoint and orthogonality for tensors. The adjoint operation corresponds to a special case of the so called contraction product [6] or n-mode product [19] in the literature, but is more intuitive to view it from an operator point of view. The first order optimality condition is then converted into a variant of the high-order power method with an added mechanism for addressing the needed orthogonality. The resulting objective values are proved to enjoy an ascending property in Section 3 and hence converge. To prove the convergence of the iterates generated by the high-order power method requires additional mathematical insights which are presented in Section 4. In particular, we relate the ALS method to a continuous projected gradient flow and argue that the stationary points are isolated.

2. Orthogonal Rank-2 approximation. The problem of finding a best rank-2 approximation to a given order-k tensor $T \in \mathbb{R}^{I_1 \times \cdots \times I_k}$ is to determine unit vectors $\mathbf{u}_1^{(\ell)}, \mathbf{u}_2^{(\ell)} \in \mathbb{R}^{I_\ell}$, $\ell = 1, \ldots k$, and scalars $\lambda_1, \lambda_2 \in \mathbb{R}$ such that the function

$$f\left(\lambda, \mu, \mathbf{u}_1^{(1)}, \ldots, \mathbf{u}_1^{(k)}, \mathbf{u}_2^{(1)}, \ldots, \mathbf{u}_2^{(k)}\right) := \left\| T - \lambda_1 \underbrace{\mathbf{u}_1^{(1)} \otimes \ldots \otimes \mathbf{u}_1^{(k)}}_{H_1} - \lambda_2 \underbrace{\mathbf{u}_2^{(1)} \otimes \ldots \otimes \mathbf{u}_2^{(k)}}_{H_2} \right\|_F^2 \tag{2.1}$$

is minimized. The complication of this seemingly straightforward least squares problem can be seen from the following example [9]. Given linearly independent unit vectors $\mathbf{a}^{(j)}, \mathbf{b}^{(j)} \in \mathbb{R}^{I_j}$, $j = 1, \ldots, k$, consider the tensor

$$X := \mathbf{b}^{(1)} \otimes \mathbf{a}^{(2)} \otimes \ldots \otimes \mathbf{a}^{(k)} + \mathbf{a}^{(1)} \otimes \mathbf{b}^{(2)} \otimes \ldots \otimes \mathbf{a}^{(k)} + \ldots + \mathbf{a}^{(1)} \otimes \mathbf{a}^{(2)} \otimes \ldots \otimes \mathbf{b}^{(k)}$$

which is of rank k. For $\gamma \in \mathbb{R}$, define

$$Y(\gamma) := \gamma \left(\mathbf{a}^{(1)} + \frac{\mathbf{b}^{(1)}}{\gamma}\right) \otimes \left(\mathbf{a}^{(2)} + \frac{\mathbf{b}^{(2)}}{\gamma}\right) \otimes \ldots \otimes \left(\mathbf{a}^{(k)} + \frac{\mathbf{b}^{(k)}}{\gamma}\right) - \gamma \mathbf{a}^{(1)} \otimes \mathbf{a}^{(2)} \otimes \ldots \otimes \mathbf{a}^{(k)}$$

which is of rank 2. Then it is easy to see that $Y(\gamma) \to X$ as $\gamma \to \infty$, showing that the best rank-2 approximation to X does not exist. What has happened is that if we rewrite the difference $\|X - Y(\gamma)\|_F^2$ in the form of (2.1) by enforcing unit base tensors H_1 and H_2, then the two corresponding scalars λ_1 and λ_2 are necessarily going to infinity in opposite sign [9, Proposition 4.8], resulting in longitudinal cancelation and ultimate convergence to X. In this case, the two unit base tensors H_1 and H_2 eventually become identical.

One possible remedy to avoid such a coalescing phenomenon is to impose the condition that the base tensors H_1 and H_2 remain orthogonal to each other, which will be assumed henceforth. This requirement is

more in line with, but not exactly the same as, the best rank-2 approximation to a matrix, where the two rank-1 base matrices formed from singular vectors are necessarily orthogonal to each other. Note that

$$\langle H_1, H_2\rangle = \prod_{\ell=1}^{k} \langle \mathbf{u}_1^{(\ell)}, \mathbf{u}_2^{(\ell)}\rangle.$$

So H_1 is orthogonal to H_2 if and only if at least one pair of components $\left(\mathbf{u}_1^{(\ell)}, \mathbf{u}_2^{(\ell)}\right)$ are orthogonal. Such an implication that only one pair of orthogonality among components of rank-1 matrices is needed is in sharp contrast to and more general than the pairwise orthogonality happened in the singular value decomposition for matrices. This notion of H_1 being orthogonal to H_2 corresponds to the so called orthogonal rank approximation which is the most general case among three definitions of orthogonality for two rank-1 tensors characterized in [17]. The more restricted case of the so called complete orthogonality requires pairwise orthogonality of all components of rank-1 matrices, of which an example is the singular value decomposition, and is different from our consideration. See [36, Section 3] for a theoretical discussion on the completely orthogonal decomposition and [17] for a general treatment of orthogonality. See also [3] and [24] for some algorithms on the completely orthogonal approximation.

The tensors H_1 and H_2 are of unit length if all vectors $\mathbf{u}_1^{(\ell)}, \mathbf{u}_2^{(\ell)}$, $\ell = 1, \ldots, k$, are of unit length. For each fixed pair of mutual orthogonal unit tensors H_1 and H_2, the optimal scalars in (2.1) are necessarily given by

$$\lambda_i = \lambda_i\left(\mathbf{u}_i^{(1)}, \ldots, \mathbf{u}_i^{(k)}\right) = \left\langle T, \mathbf{u}_i^{(1)} \otimes \ldots \otimes \mathbf{u}_i^{(k)} \right\rangle, \quad i = 1, 2. \tag{2.2}$$

Note that the orthogonality effectively implies that $|\lambda_i| \leq \|T\|_F$, making a difference to that without the orthogonality. To put it differently, the expression for λ_i in (2.2) is precisely the length of the projection of the "vector" T onto the the "unit vector" H_i under the Frobenius inner product for tensors. Thus, the minimization of (2.1) is a least squares problem and is equivalent to

$$\max_{\|H_1\|_F = \|H_2\|_F = 1, \langle H_1, H_2\rangle = 0} \quad \lambda_1^2 + \lambda_2^2. \tag{2.3}$$

For simplicity, we shall relax the condition of orthogonality $\langle H_1, H_2\rangle = 0$ to $\langle \mathbf{u}_1^{(s)}, \mathbf{u}_2^{(s)}\rangle = 0$ for an indeterminate integer s between 1 and k.

2.1. Adjoint operation. For the simplicity of discussion, we introduce a generalized notion of adjoint. We begin with an order-3 tensor T so as to visualize it as a $p \times 1$ block matrix of which each block is an $m \times n$ matrices, denoted by $\tau_{i,:}$, whereas the linear operator $\mathscr{T} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ represented by T has the action $\circledast$ defined by

$$\mathscr{T}(X) \equiv \underbrace{\begin{bmatrix} \boxed{\tau_{1,:}} \\ \boxed{\tau_{2,:}} \\ \vdots \\ \boxed{\tau_{p,:}} \end{bmatrix}}_{T} \circledast X := [\langle \tau_{i,:}, X\rangle] \in \mathbb{R}^p.$$

In this way, the adjoint operator $\mathscr{T}^* : \mathbb{R}^p \to \mathbb{R}^{m\times n}$ should be defined as

$$\mathscr{T}^*(\mathbf{v}) := \sum_i v_i \tau_{i,:} \tag{2.4}$$

which satisfies the requirement that

$$\langle \mathscr{T}(X), \mathbf{v}\rangle = \langle X, \mathscr{T}^*(\mathbf{v})\rangle \tag{2.5}$$

for every $X \in \mathbb{R}^{m\times n}$ and $\mathbf{v} \in \mathbb{R}^p$.

A similar idea can be generalized to an order-k tensor $T \in \mathbb{R}^{I_1 \times \ldots \times I_k}$ by identifying T as the operator $\mathscr{T} \equiv \mathscr{T}_\ell$, $\ell = 1, \ldots, k$, with

$$\mathscr{T}_\ell : \mathbb{R}^{I_1 \times \ldots \times \widehat{I}_\ell \times \ldots \times I_k} \to \mathbb{R}^{I_\ell}, \tag{2.6}$$

where $\widehat{I}_\ell$ means that quantities associated with this particular index ℓ are taken out from the remaining list, and defining its action at any $S \in \mathbb{R}^{I_1 \times \ldots \times \widehat{I}_\ell \times \ldots \times I_k}$ by the operation

$$\mathscr{T}_\ell(S) := T \circledast_\ell S = [\langle \tau_{:,i_\ell,:}, S \rangle] \in \mathbb{R}^{I_\ell}, \tag{2.7}$$

where $\tau_{:,i_\ell,:}$ denotes the i_ℓ-th "slice" of the tensor T in the ℓ-th direction and $\langle \cdot, \cdot \rangle$ is the Frobenius inner product generalized to multi-dimensional arrays. Trivially, the real-valued order-k tensors form a vector space equipped with the Frobenius inner product. We may therefore postulate the idea of the notion of $\mathscr{T}^*$. The following result generalizes the spirit of (2.5), whereas the notion of "adjoint" is embedded in the tensor product $\circledast_\ell$. Its proof is only a rearrangement of terms in the summation by the associative law.

LEMMA 2.1. *For any $\ell = 1, \ldots, k$, let $\mathbf{h}^{(\ell)} \in \mathbb{R}^{I_\ell}$ be an arbitrary vector. Then it is true that*

$$\left\langle T, \mathbf{u}^{(1)} \otimes \ldots \otimes \mathbf{h}^{(\ell)} \otimes \ldots \otimes \mathbf{u}^{(k)} \right\rangle = \left\langle T \circledast_\ell \left(\mathbf{u}^{(1)} \otimes \ldots \otimes \widehat{\mathbf{u}}^{(\ell)} \otimes \ldots \otimes \mathbf{u}^{(k)} \right), \mathbf{h}^{(\ell)} \right\rangle. \tag{2.8}$$

Because (2.8) holds for any index ℓ, the reciprocating relationship

$$\left\langle T \circledast_i \left(\mathbf{u}^{(1)} \otimes \ldots \otimes \widehat{\mathbf{u}}^{(i)} \otimes \ldots \otimes \mathbf{u}^{(k)} \right), \mathbf{u}^{(i)} \right\rangle = \left\langle T \circledast_j \left(\mathbf{u}^{(1)} \otimes \ldots \otimes \widehat{\mathbf{u}}^{(j)} \otimes \ldots \otimes \mathbf{u}^{(k)} \right), \mathbf{u}^{(j)} \right\rangle. \tag{2.9}$$

then holds for any two distinct indices i and j. Both identities (2.8) and (2.9) will be extremely useful in the subsequent discussion.

2.2. Simultaneous subspace iteration. Returning to the constrained optimization problem (2.3), we formulate its Lagrangian $\mathcal{L} = \mathcal{L}\left(\mathbf{u}_1^{(1)}, \ldots, \mathbf{u}_1^{(k)}, \mathbf{u}_2^{(1)}, \ldots, \mathbf{u}_2^{(k)}, \rho_1^{(1)}, \ldots, \rho_1^{(k)}, \rho_2^{(1)}, \ldots, \rho_2^{(k)}, \mu\right)$ as

$$\mathcal{L} := \lambda_1^2 + \lambda_2^2 - \sum_{\ell=1}^{k} \left(\rho_1^{(\ell)} \left(\langle \mathbf{u}_1^{(\ell)}, \mathbf{u}_1^{(\ell)} \rangle - 1 \right) + \rho_2^{(\ell)} \left(\langle \mathbf{u}_2^{(\ell)}, \mathbf{u}_2^{(\ell)} \rangle - 1 \right) \right) - \mu \langle \mathbf{u}_1^{(s)}, \mathbf{u}_2^{(s)} \rangle.$$

Accordingly, the first order optimality condition for a stationary point is to satisfy the system of equations defined by [22]

$$\lambda_i T \circledast_\ell \left(\mathbf{u}_i^{(1)} \otimes \ldots \otimes \widehat{\mathbf{u}}_i^{(\ell)} \otimes \ldots \otimes \mathbf{u}_i^{(k)} \right) = \rho_i^{(\ell)} \mathbf{u}_i^{(\ell)}, \quad \ell = 1, \ldots, \widehat{s}, \ldots, k, \quad i = 1, 2, \tag{2.10}$$

and

$$\begin{cases} \lambda_1 \underbrace{T \circledast_s \left(\mathbf{u}_1^{(1)} \otimes \ldots \otimes \widehat{\mathbf{u}}_1^{(s)} \otimes \ldots \otimes \mathbf{u}_1^{(k)} \right)}_{\mathbf{v}_1^{(s)}} & = & \rho_1^{(s)} \mathbf{u}_1^{(s)} + \frac{\mu}{2} \mathbf{u}_2^{(s)}, \\ \lambda_2 \underbrace{T \circledast_s \left(\mathbf{u}_2^{(1)} \otimes \ldots \otimes \widehat{\mathbf{u}}_2^{(s)} \otimes \ldots \otimes \mathbf{u}_2^{(k)} \right)}_{\mathbf{v}_2^{(s)}} & = & \rho_2^{(s)} \mathbf{u}_2^{(s)} + \frac{\mu}{2} \mathbf{u}_1^{(s)}. \end{cases} \tag{2.11}$$

It is easy to see that the Lagrange multipliers satisfy the relationships

$$\rho_i^{(\ell)} = \lambda_i^2, \quad \ell = 1, \ldots, k, \tag{2.12}$$

and

$$\frac{\mu}{2} = \lambda_1 \langle \mathbf{v}_1^{(s)}, \mathbf{u}_2^{(s)} \rangle = \lambda_2 \langle \mathbf{v}_2^{(s)}, \mathbf{u}_1^{(s)} \rangle, \tag{2.13}$$

which then reduces the systems (2.10) to

$$T\circledast_\ell \left(\mathbf{u}_i^{(1)}\otimes\ldots\otimes\widehat{\mathbf{u}}_i^{(\ell)}\otimes\ldots\otimes\mathbf{u}_i^{(k)}\right)=\lambda_i\mathbf{u}_i^{(\ell)},\quad \ell=1,\ldots,\widehat{s},\ldots,k,\quad i=1,2, \tag{2.14}$$

and (2.11) to

$$\left\{\begin{array}{rcl}\mathbf{v}_1^{(s)} &=& \lambda_1\mathbf{u}_1^{(s)}+\langle\mathbf{v}_1^{(s)},\mathbf{u}_2^{(s)}\rangle\mathbf{u}_2^{(s)},\\ \mathbf{v}_2^{(s)} &=& \lambda_2\mathbf{u}_2^{(s)}+\langle\mathbf{v}_2^{(s)},\mathbf{u}_1^{(s)}\rangle\mathbf{u}_1^{(s)}.\end{array}\right. \tag{2.15}$$

In all, (2.14) and (2.15) together represent a system of nonlinear equations for the unknowns $\mathbf{u}_1^{(1)},\ldots,\mathbf{u}_1^{(k)}$ and $\mathbf{u}_2^{(1)},\ldots,\mathbf{u}_2^{(k)}$. A solution to this system gives rise to a stationary point of (2.3).

A commonly used tactic is to alternate directions among the variables in a way similar to the classical Gauss-Seidel method. This is also known as the high-order power method. Our added challenge is to maintain the conditions of orthogonality, which we handle as follows.

Denote $\xi_i:=\frac{\mu}{2\lambda_i}$, $i=1,2$. Then a "temporary" solution to (2.15) can be expressed as

$$\left\{\begin{array}{rcl}\mathbf{u}_1^{(s)} &=& \frac{1}{\lambda_1\lambda_2-\xi_1\xi_2}\Big(\lambda_2\mathbf{v}_1^{(s)}-\xi_1\mathbf{v}_2^{(s)}\Big),\\ \mathbf{u}_2^{(s)} &=& \frac{1}{\lambda_1\lambda_2-\xi_1\xi_2}\Big(\lambda_1\mathbf{v}_2^{(s)}-\xi_2\mathbf{v}_1^{(s)}\Big).\end{array}\right. \tag{2.16}$$

We say that (2.16) is temporary because ξ_1 and ξ_2 on the right-hand side also depend on $\mathbf{u}_1^{(s)}$ and $\mathbf{u}_2^{(s)}$ implicitly. Using the fact that $\mathbf{u}_1^{(s)}$ is orthogonal to $\mathbf{u}_2^{(s)}$, we find that

$$\langle\mathbf{v}_1^{(s)},\mathbf{v}_1^{(s)}\rangle+\langle\mathbf{v}_2^{(s)},\mathbf{v}_2^{(s)}\rangle=\left(\frac{\lambda_2}{\xi_1}+\frac{\xi_1}{\lambda_2}\right)\langle\mathbf{v}_1^{(s)},\mathbf{v}_2^{(s)}\rangle.$$

It follows that

$$t:=\frac{\xi_1}{\lambda_2}=\frac{\xi_2}{\lambda_1}=\frac{c\pm\sqrt{c^2-4}}{2}, \tag{2.17}$$

where, assuming that $\langle\mathbf{v}_1^{(s)},\mathbf{v}_2^{(s)}\rangle\neq 0$, the scalar

$$c:=\frac{\langle\mathbf{v}_1^{(s)},\mathbf{v}_1^{(s)}\rangle+\langle\mathbf{v}_2^{(s)},\mathbf{v}_2^{(s)}\rangle}{\langle\mathbf{v}_1^{(s)},\mathbf{v}_2^{(s)}\rangle} \tag{2.18}$$

has modulus $|c|\geq 2$ and, hence, t is real. There are two choices of t to satisfy the necessary condition (2.11). We shall use the parameter t to maintain the orthogonality and to control the objective value.

We propose an iterative procedure as is described in Algorithm 1. For clarity, we have included the subscript $\cdot_{i,[p]}$ to indicate that the associated quantity is being resulted from the p-th iteration, though in a real coding such a reference is not needed. It is important to note that between the two possible values of t specified in (2.17), the final choice of t selected in the algorithm depends on the sign of $\langle\mathbf{v}_1^{(s)},\mathbf{v}_2^{(s)}\rangle$. The purpose is to introduce a smaller additive perturbation so as to maintain a certain monotonicity which will be explained in the next section.

3. Convergence of objective values. The iterates defined by Algorithm 1 actually enjoy an inherent ascending property. In this section, we explain how they can be regarded as a variant of the so called alternative least squares method favored in the literature [1, 12]. As the expressions become repetitive and tedious, we shall adopt the following abbreviations for convenience. For integers $\alpha\leq\beta$, let

$$U_{i,[p]}^{\{\alpha..\beta\}}:=\left(\mathbf{u}_{i,[p]}^{(\alpha)},\mathbf{u}_{i,[p]}^{(\alpha+1)},\ldots,\mathbf{u}_{i,[p]}^{(\beta)}\right)\in\mathbb{R}^{I_\alpha}\times\ldots\times\mathbb{R}^{I_\beta} \tag{3.1}$$

Algorithm 1 (High-order power method for orthogonal rank decomposition.)

Given $\mathbf{u}_{1,[0]}^{(1)}, \mathbf{u}_{2,[0]}^{(1)} \in \mathbb{R}^{I_1}, \ldots, \mathbf{u}_{1,[0]}^{(k)}, \mathbf{u}_{2,[0]}^{(k)} \in \mathbb{R}^{I_k}$, and $\mathbf{u}_{1,[0]}^{(s)} \perp \mathbf{u}_{2,[0]}^{(s)}$,
for $p = 0, 1, \cdots,$ **do**
  for $\ell = 1, 2, \cdots, k$ **do**
    if $\ell = s$ **then**
      $\mathbf{v}_{1,[p+1]}^{(s)} = T \circledast_\ell \left(\mathbf{u}_{1,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{1,[p+1]}^{(s-1)} \otimes \widehat{\mathbf{u}}^{(s)} \otimes \mathbf{u}_{1,[p]}^{(s+1)} \ldots \otimes \mathbf{u}_{1,[p]}^{(k)} \right)$
      $\mathbf{v}_{2,[p+1]}^{(s)} = T \circledast_\ell \left(\mathbf{u}_{2,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{2,[p+1]}^{(s-1)} \otimes \widehat{\mathbf{u}}^{(s)} \otimes \mathbf{u}_{2,[p]}^{(s+1)} \ldots \otimes \mathbf{u}_{2,[p]}^{(k)} \right)$
      if $\langle \mathbf{v}_{1,[p+1]}^{(s)}, \mathbf{v}_{2,[p+1]}^{(s)} \rangle = 0$ **then**
        $t = 0$
      else
        $c := \frac{\langle \mathbf{v}_{1,[p+1]}^{(s)}, \mathbf{v}_{1,[p+1]}^{(s)} \rangle + \langle \mathbf{v}_{2,[p+1]}^{(s)}, \mathbf{v}_{2,[p+1]}^{(s)} \rangle}{\langle \mathbf{v}_{1,[p+1]}^{(s)}, \mathbf{v}_{2,[p+1]}^{(s)} \rangle}$
        if $c > 0$ **then**
          $t = \frac{c - \sqrt{c^2 - 4}}{2}$
        else if $c < 0$ **then**
          $t = \frac{c + \sqrt{c^2 - 4}}{2}$
        end if
      end if
      $\mathbf{u}_{1,[p+1]}^{(s)} = \mathbf{v}_{1,[p+1]}^{(s)} - t \mathbf{v}_{2,[p+1]}^{(s)}$
      $\mathbf{u}_{2,[p+1]}^{(s)} = \mathbf{v}_{2,[p+1]}^{(s)} - t \mathbf{v}_{1,[p+1]}^{(s)}$
      for $i = 1, 2$ **do**
        $\lambda_{i,[p+1]}^{(s)} := \| \mathbf{u}_{i,[p+1]}^{(s)} \|_2$
        $\mathbf{u}_{i,[p+1]}^{(s)} := \frac{\mathbf{u}_{i,[p+1]}^{(s)}}{\lambda_{i,[p+1]}^{(s)}}$
      end for
    else
      for $i = 1, 2$ **do**
        $\mathbf{u}_{i,[p+1]}^{(\ell)} = T \circledast_\ell \left(\mathbf{u}_{i,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{i,[p+1]}^{(\ell-1)} \otimes \widehat{\mathbf{u}}^{(\ell)} \otimes \mathbf{u}_{i,[p]}^{(\ell+1)} \ldots \otimes \mathbf{u}_{i,[p]}^{(k)} \right)$
        $\lambda_{i,[p+1]}^{(\ell)} := \| \mathbf{u}_{i,[p+1]}^{(\ell)} \|_2$
        $\mathbf{u}_{i,[p+1]}^{(\ell)} := \frac{\mathbf{u}_{i,[p+1]}^{(\ell)}}{\lambda_{i,[p+1]}^{(\ell)}}$
      end for
    end if
  end for
end for

denote the list of vectors $\mathbf{u}_{i,[p]}^{(\ell)}$, $\ell = \alpha, \ldots, \beta$, and

$$\bigotimes U_{i,[p]}^{\{\alpha..\beta\}} := \mathbf{u}_{i,[p]}^{(\alpha)} \otimes \mathbf{u}_{i,[p]}^{(\alpha+1)} \otimes \ldots \otimes \mathbf{u}_{i,[p]}^{(\beta)} \tag{3.2}$$

the outer product of these vectors.

For each fixed p and ℓ, suppose vectors $\mathbf{u}_{i,[p+1]}^{(j)}$, $j = 1, \ldots, \ell - 1$ and $\mathbf{u}_{i,[p]}^{(j)}$, $j = \ell + 1, \ldots, k$ are given. Consider the maximization of the functional $\theta_\ell : \mathbb{R}^{I_\ell} \to \mathbb{R}$ defined by

$$\theta^{(\ell)} \left(\mathbf{y}_1^{(\ell)}, \mathbf{y}_2^{(\ell)} \right) := \sum_{i=1}^{2} \left\langle T, \bigotimes U_{i,[p+1]}^{\{1..(\ell-1)\}} \otimes \mathbf{y}_i^{(\ell)} \otimes \bigotimes U_{i,[p]}^{\{(\ell+1)..k\}} \right\rangle^2, \quad \ell = 1, \ldots k, \tag{3.3}$$

subject to the constraints $\|\mathbf{y}_i^{(\ell)}\|_2 = 1$ and $\left\langle \mathbf{y}_1^{(s)}, \mathbf{y}_2^{(s)} \right\rangle = 0$ when $\ell = s$. For convenience, write

$$\pi_{i,[p+1]}^{(\ell)} = \pi_{i,[p+1]}^{(\ell)}\left(\mathbf{y}_i^{(\ell)}\right) := \left\langle T, \bigotimes U_{i,[p+1]}^{\{1..(\ell-1)\}} \otimes \mathbf{y}_i^{(\ell)} \otimes \bigotimes U_{i,[p]}^{\{(\ell+1)..k\}} \right\rangle. \tag{3.4}$$

Applying the Lagrange multiplier theory to (3.3) and using (2.8), it is easy to see that the first order optimality condition for (3.3) are the equations

$$\pi_{i,[p+1]}^{(\ell)} T \circledast_\ell \left(\bigotimes U_{i,[p+1]}^{\{1..(\ell-1)\}} \otimes \widehat{\mathbf{y}}_i^{(\ell)} \otimes \bigotimes U_{i,[p]}^{\{(\ell+1)..k\}} \right) = \eta_i^{(\ell)} \mathbf{y}_i^{(\ell)}, \tag{3.5}$$

for $\ell = 1, \ldots, \widehat{s}, \ldots, k$, $i = 1, 2$; and

$$\begin{cases} \pi_{1,[p+1]}^{(s)} \underbrace{T \circledast_s \left(\bigotimes U_{1,[p+1]}^{\{(1..(s-1))\}} \otimes \widehat{\mathbf{y}}_1^{(s)} \otimes \bigotimes U_{1,[p]}^{\{((s+1)..k)\}} \right)}_{\mathbf{v}_{1,[p+1]}^{(s)}} &= \eta_1^{(s)} \mathbf{y}_1^{(s)} + \frac{\mu^{(s)}}{2} \mathbf{y}_2^{(s)}, \\ \pi_{2,[p+1]}^{(s)} \underbrace{T \circledast_s \left(\bigotimes U_{2,[p+1]}^{\{(1..(s-1))\}} \otimes \widehat{\mathbf{y}}_2^{(s)} \otimes \bigotimes U_{2,[p]}^{\{((s+1)..k)\}} \right)}_{\mathbf{v}_{2,[p+1]}^{(s)}} &= \eta_2^{(s)} \mathbf{y}_2^{(s)} + \frac{\mu^{(s)}}{2} \mathbf{y}_1^{(s)}, \end{cases} \tag{3.6}$$

where $\eta_i^{(\ell)} = \left(\pi_{i,[p+1]}^{(\ell)}\right)^2 > 0$ and $\mu^{(s)}$ are just some Lagrangian multipliers. For $\ell = 1, \ldots, \widehat{s}, \ldots, k$ and $i = 1, 2$, it follows that

$$\left|\pi_{i,[p+1]}^{(\ell)}\right| = \left\| T \circledast_\ell \left(\bigotimes U_{i,[p+1]}^{\{(1..(\ell-1))\}} \otimes \widehat{\mathbf{y}}_i^{(\ell)} \otimes \bigotimes U_{i,[p]}^{\{(\ell+1)..k)\}} \right) \right\| = \lambda_{i,[p+1]}^{(\ell)}. \tag{3.7}$$

By flipping $\mathbf{y}_i^{(\ell)}$ to $-\mathbf{y}_i^{(\ell)}$, if necessary, we may assume without loss of generality that $\pi_{i,[p+1]}^{(\ell)} > 0$. In this way, since the vector $T \circledast_\ell \left(\mathbf{u}_{i,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{i,[p+1]}^{(\ell-1)} \otimes \widehat{\mathbf{y}}_i^{(\ell)} \otimes \mathbf{u}_{i,[p]}^{(\ell+1)} \otimes \ldots \otimes \mathbf{u}_{i,[p]}^{(k)} \right)$ points to an ascent direction for $\theta^{(\ell)}$, we conclude that the pair of vectors $\left(\mathbf{u}_{1,[p+1]}^{(\ell)}, \mathbf{u}_{2,[p+1]}^{(\ell)} \right)$ defined in Algorithm 1 is a local maximizer for the functionals $\theta^{(\ell)}$, $\ell = 1, \ldots \widehat{s}, \ldots k$.

As for the case $\ell = s$, we have

$$\begin{cases} \mathbf{v}_{1,[p+1]}^{(s)} &= \pi_{1,[p+1]} \mathbf{y}_1^{(s)} + \langle \mathbf{v}_{1,[p+1]}^{(s)}, \mathbf{y}_2^{(s)} \rangle \mathbf{y}_2^{(s)}, \\ \mathbf{v}_{2,[p+1]}^{(s)} &= \pi_{2,[p+1]} \mathbf{y}_2^{(s)} + \langle \mathbf{v}_{2,[p+1]}^{(s)}, \mathbf{y}_1^{(s)} \rangle \mathbf{y}_1^{(s)}. \end{cases} \tag{3.8}$$

which is similar to the system (2.15). Thus, by employing the same techniques, we may derive the pair of vectors $\left\{ \mathbf{u}_{1,[p+1]}^{(s)}, \mathbf{u}_{2,[p+1]}^{(s)} \right\}$ with the choice of c and t characterized in Algorithm 1. It can be verified that by the particular choice of t specified in Algorithm1 we always have

$$\begin{cases} \left\langle \mathbf{u}_{1,[p+1]}^{(s)}, \mathbf{v}_{1,[p+1]}^{(s)} \right\rangle = \frac{1}{2\lambda_{1,[p+1]}^{(s)}} \left(\|\mathbf{v}_{1,[p+1]}^{(s)}\|^2 - \|\mathbf{v}_{2,[p+1]}^{(s)}\|^2 + \|\mathbf{v}_{1,[p+1]}^{(s)} - \mathbf{v}_{2,[p+1]}^{(s)}\| \|\mathbf{v}_{1,[p+1]}^{(s)} + \mathbf{v}_{2,[p+1]}^{(s)}\| \right) \geq 0, \\ \left\langle \mathbf{u}_{2,[p+1]}^{(s)}, \mathbf{v}_{1,[p+1]}^{(s)} \right\rangle = \frac{1}{2\lambda_{2,[p+1]}^{(s)}} \left(\|\mathbf{v}_{2,[p+1]}^{(s)}\|^2 - \|\mathbf{v}_{1,[p+1]}^{(s)}\|^2 + \|\mathbf{v}_{1,[p+1]}^{(s)} - \mathbf{v}_{2,[p+1]}^{(s)}\| \|\mathbf{v}_{1,[p+1]}^{(s)} + \mathbf{v}_{2,[p+1]}^{(s)}\| \right) \geq 0. \end{cases}$$

That is, the two vectors $\mathbf{u}_{i,[p+1]}^{(2)}$ and $\mathbf{v}_{i,[p+1]}^{(s)}$ maintain an acute angle for both $i = 1, 2$, respectively. Since $\mathbf{v}_{i,[p+1]}^{(s)}$ points to the gradient direction of $\theta^{(s)}$, the choice of $\mathbf{u}_{i,[p+1]}^{(s)}$ is the only other unit vector that is nearest to $\mathbf{v}_i^{(s)}$ and satisfies the optimality condition. Thus, the pair of vectors $\left(\mathbf{u}_{1,[p+1]}^{(s)}, \mathbf{u}_{2,[p+1]}^{(s)} \right)$ is a local maximizer

for the functionals $\theta^{(s)}$, returning an objective value better than that of the pair $\left\{\mathbf{u}_{1,[p]}^{(s)}, \mathbf{u}_{2,[p]}^{(s)}\right\}$ which satisfies only the condition of orthogonality.

All together, we have shown the following ascent property.

LEMMA 3.1. *For each fixed p and ℓ, the pair of vectors $\left(\mathbf{u}_{1,[p+1]}^{(\ell)}, \mathbf{u}_{2,[p+1]}^{(\ell)}\right)$ defined by Algorithm 1 forms a pair of maximizers for the functional $\theta^{(\ell)}$ defined in (3.3). Additionally, the inequality*

$$\theta^{(\ell-1)}\left(\mathbf{u}_{1,[p+1]}^{(\ell-1)}, \mathbf{u}_{2,[p+1]}^{(\ell-1)}\right) \leq \theta^{(\ell)}\left(\mathbf{u}_{1,[p+1]}^{(\ell)}, \mathbf{u}_{2,[p+1]}^{(\ell)}\right), \quad \ell = 2, \ldots, k,$$

holds, including the case when $\ell = s$.

By repeating the above argument to successive ℓ, Algorithm 1 works exactly like the mechanism adopted in the classical Gauss-Seidel method. We alter one pair of variables at a time while keeping all other variables unchanged and gradually migrate from the set of variables from $U_{i,[p]}^{\{1..k\}}$ to $U_{i,[p+1]}^{\{1..k\}}$. During this process of transition, the intermediate values $\sum_{i=1}^{2}\left\langle T, \bigotimes U_{i,[p+1]}^{\{1..(\ell-1)\}} \otimes \mathbf{u}_{i,[p+1]}^{(\ell)} \otimes \bigotimes U_{i,[p]}^{\{(\ell+1)..k\}}\right\rangle^2$ are also moving upward, but are sandwiched in between the two extremes,

$$\sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{i,[p]}^{(k)}\right\rangle^2 \leq \sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{i,[p+1]}^{(k)}\right\rangle^2. \tag{3.9}$$

In other words, each sweep in the p loop of Algorithm 1 has the effect of increasing the objective value of $\lambda_1^2 + \lambda_2^2$. Because the range of λ_i is bounded by $\|T\|_F$, the sequence $\left\{\sum_{i=1}^{2}\lambda_i^2\left(\mathbf{u}_{i,[p]}^{(1)}, \ldots, \mathbf{u}_{i,[p]}^{(k)}\right)\right\}$ of objective values generated from iterates by Algorithm 1 must converge.

4. Convergence of iterates. It is known that the iterates from a general ALS method may not converge even to a stationary point [19]. The argument in the preceding section shows only the nondecreasing property of the objective values $\lambda_1^2 + \lambda_2^2$, which is not enough yet to guarantee the convergence of iterates themselves to a minimizer. More need be done. For any fixed p, define the incremental quantities

$$\delta\mathbf{u}_{i,[p]}^{(\ell)} := \mathbf{u}_{i,[p+1]}^{(\ell)} - \mathbf{u}_{i,[p]}^{(\ell)}, \quad \ell = 1, \ldots, k, \quad , i = 1, 2. \tag{4.1}$$

We are interested in relating the size of $\left\|\delta\mathbf{u}_{i,[p]}^{(\ell)}\right\|$ to the variation of the objective values and use this estimate to prove the converges of the iterates.

4.1. Preparatory lemmas. To show that $\delta\mathbf{u}_{i,[p]}^{(\ell)}$ diminishes to zero, we shall make use the following simple fact in subsequent development [33, 35].

LEMMA 4.1. *Suppose that $\mathbf{u}_{i,[p+1]}^{(\ell)}$ and $\mathbf{u}_{i,[p]}^{(\ell)}$ are unit vectors. Then*

$$\left\langle \delta\mathbf{u}_{i,[p]}^{(\ell)}, \delta\mathbf{u}_{i,[p]}^{(\ell)}\right\rangle = 2\left\langle \mathbf{u}_{i,[p+1]}^{(\ell)}, \delta\mathbf{u}_{i,[p]}^{(\ell)}\right\rangle. \tag{4.2}$$

Proof. By the definition (4.1), it is trivial to see that $\left\langle \delta\mathbf{u}_{i,[p]}^{(\ell)}, \delta\mathbf{u}_{i,[p]}^{(\ell)}\right\rangle = 2 - 2\left\langle \mathbf{u}_{i,[p+1]}^{(\ell)}, \mathbf{u}_{i,[p]}^{(\ell)}\right\rangle$. The assertion follows from $\left\langle \mathbf{u}_{i,[p+1]}^{(\ell)}, \delta\mathbf{u}_{i,[p]}^{(\ell)}\right\rangle = \left\langle \mathbf{u}_{i,[p+1]}^{(\ell)}, \mathbf{u}_{i,[p+1]}^{(\ell)} - \mathbf{u}_{i,[p]}^{(\ell)}\right\rangle = 1 - \left\langle \mathbf{u}_{i,[p+1]}^{(\ell)}, \mathbf{u}_{i,[p]}^{(\ell)}\right\rangle$. □

To argue that the sequences $\left\{\mathbf{u}_{i,[p]}^{(\ell)}\right\}$ also converges, we need the next lemma.

LEMMA 4.2. *Let $\{a_n\}$ be a bounded sequence of real numbers with the property that the difference $a_{n+1} - a_n$ converges to zero as n goes to infinity. If the limit points for the sequence are isolated, then the sequence $\{a_n\}$ converges to a unique limit point.*

Proof. We shall prove by contradiction. Suppose that $\{a_{\alpha_k}\}$ and $\{a_{\beta_k}\}$ are two subsequences of $\{a_k\}$ converging, respectively, to two distinct limit points, x and y. Let z denote any fixed real number in between x and y.

Let $B_x(r)$ denote the interval $[x-r, x+r]$ centered x. For any $0 < \epsilon < \frac{1}{4}\min\{|x-z|, |y-z|\}$, there exists a large enough integer $K = K(\epsilon)$ such that

$$\left\{ \begin{array}{rcl} a_{\alpha_k} & \in & B_x(\epsilon) \\ a_{\beta_k} & \in & B_y(\epsilon) \\ |a_{k+1} - a_k| & < & \epsilon \end{array} \right.$$

for all $k \geq K$. But infinitely many elements of $\{a_k\}$ must leave $B_x(\epsilon)$ to enter $B_y(\epsilon)$ and vise versa, implying that there are infinitely many elements reaching the neighborhood $B_z(\epsilon)$. It follows that z is also a limit point. Since z is arbitrary, any number in between x and y is a limit point. This contradicts the assumption that there these limit points are isolated. □

Finally, to take advantage of Lemma 4.2, we need to establish the next result showing that the limit points of Algorithm 1 are necessarily the equilibria of an analytic gradient flow.

LEMMA 4.3. *The limit points of any sequence generated by Algorithm 1 are necessarily isolated. That is, around any limit point there is a neighborhood such that the sequence of iterates from Algorithm 1 has no other limit point.*

Proof. The limit points, if there is any, necessarily satisfy the system of equations (2.14) and (2.15). So it suffices to prove that this nonlinear system has only isolated solutions. Our plan of proof is to show that these solutions are stationary points of an analytic gradient flow, which then by a notion already established in the field of dynamical systems must be isolated.

Consider the objection function in (2.3)

$$\phi\left(\mathbf{u}_1^{(1)}, \mathbf{u}_1^{(2)}, \ldots, \mathbf{u}_1^{(k)}, \mathbf{u}_2^{(1)}, \mathbf{u}_2^{(2)}, \ldots, \mathbf{u}_2^{(k)}\right) := \sum_{i=1}^{2} \left\langle T, \mathbf{u}_i^{(1)} \otimes \ldots \otimes \mathbf{u}_i^{(k)} \right\rangle^2. \tag{4.3}$$

Trivially the (partial) Fréchet derivative of ϕ with respect to $\mathbf{u}_i^{(\ell)}$ acting on $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^{I_\ell}$ is given by

$$\frac{\partial \phi}{\partial \mathbf{u}_i^{(\ell)}}.\mathbf{h}_i^{(\ell)} = 2\lambda_i \left\langle T, \mathbf{u}_i^{(1)} \otimes \mathbf{u}^{(2)} \otimes \ldots \otimes \mathbf{h}_i^{(\ell)} \otimes \ldots \otimes \mathbf{u}^{(k)} \right\rangle. \tag{4.4}$$

So by the Riesz representation theorem, we obtain

$$\nabla_i^{(\ell)} \phi := 2\lambda_i T \circledast_\ell \left(\mathbf{u}_i^{(1)} \otimes \mathbf{u}_i^{(2)} \otimes \ldots \otimes \widehat{\mathbf{h}}^{(\ell)} \otimes \ldots \otimes \mathbf{u}_i^{(k)}\right) \tag{4.5}$$

as the (partial) gradient of ϕ with respect to the variable $\mathbf{u}_i^{(\ell)}$.

For the objective function to ascend continuously while satisfying the constraints, we need to project the gradient onto the tangent space of the feasible set. For $\ell \neq s$, the constraint is merely the unit ball $\|\mathbf{u}_i^{(\ell)}\| = 1$. So the projected gradient is precisely given by

$$\mathbf{Proj}_{\|\mathbf{u}_i^{(\ell)}\|=1} \nabla_i^{(\ell)} \phi = \nabla_i^{(\ell)} \phi - \left\langle \nabla_i^{(\ell)} \phi, \mathbf{u}_i^{(\ell)} \right\rangle \mathbf{u}_i^{(\ell)}. \tag{4.6}$$

The dynamical system

$$\frac{d\mathbf{u}_i^{(\ell)}}{dt} = T \circledast_\ell \left(\mathbf{u}_i^{(1)} \otimes \mathbf{u}_i^{(2)} \otimes \ldots \otimes \widehat{\mathbf{h}}^{(\ell)} \otimes \ldots \otimes \mathbf{u}_i^{(k)}\right) - \lambda_i \mathbf{u}_i^{(\ell)}, \quad \ell = 1, \ldots, \widehat{s}, \ldots, k, \quad , i = 1, 2 \tag{4.7}$$

therefore defines a gradient flow $\mathbf{u}_i^{(\ell)}(t)$ on the unit ball for the function ϕ. When $\ell = s$, we need to consider the pair $\left(\mathbf{u}_1^{(s)}, \mathbf{u}_2^{(s)}\right)$ together. It is easy to see that the projector onto the tangent space of their feasible set is precisely the $2I_s \times 2I_s$ matrix

$$\mathcal{P} := I - \left[\begin{array}{cc} \mathbf{u}_1^{(s)} \mathbf{u}_1^{(s)^\top} + \frac{1}{2} \mathbf{u}_2^{(s)} \mathbf{u}_2^{(s)^\top} & \frac{1}{2} \mathbf{u}_2^{(s)} \mathbf{u}_1^{(s)^\top} \\ \frac{1}{2} \mathbf{u}_1^{(s)} \mathbf{u}_2^{(s)^\top} & \mathbf{u}_2^{(s)} \mathbf{u}_2^{(s)^\top} + \frac{1}{2} \mathbf{u}_1^{(s)} \mathbf{u}_1^{(s)^\top} \end{array} \right]. \tag{4.8}$$

Thus the dynamical system for the project gradient flow $\left(\mathbf{u}_1^{(s)}(t), \mathbf{u}_2^{(s)}(t)\right)$ is defined by

$$\frac{d}{dt}\begin{bmatrix} \mathbf{u}_1^{(s)} \\ \mathbf{u}_2^{(s)} \end{bmatrix} = \mathcal{P}\begin{bmatrix} \nabla_1^{(s)}\phi \\ \nabla_2^{(s)}\phi \end{bmatrix} = \begin{bmatrix} 2\lambda_1\nabla_1^{(s)}\phi - 2\varrho_1^{(s)}\mathbf{u}_1^{(s)} - \nu\mathbf{u}_2^{(s)} \\ 2\lambda_2\nabla_2^{(s)}\phi - 2\varrho_2^{(s)}\mathbf{u}_2^{(s)} - \nu\mathbf{u}_1^{(s)} \end{bmatrix} \tag{4.9}$$

where

$$\begin{cases} \varrho_1^{(s)} & := & \lambda_1 {\mathbf{u}_1^{(s)}}^\top \nabla_1^{(s)}\phi, \\ \varrho_2^{(s)} & := & \lambda_2 {\mathbf{u}_2^{(s)}}^\top \nabla_2^{(s)}\phi, \\ \nu & := & \lambda_1 {\mathbf{u}_2^{(s)}}^\top \nabla_1^{(s)}\phi + \lambda_2 {\mathbf{u}_1^{(s)}}^\top \nabla_2^{(s)}\phi. \end{cases} \tag{4.10}$$

It is obvious that the stationary points of the flow (4.7) satisfy precisely the equation (2.10). A close examination of the parameters defined in (4.10) shows that $\varrho_i^{(s)}$ plays the role of $\rho_i^{(s)}$ and ν plays μ in (2.11). As such, the stationary points of the flow (4.9) satisfy precisely the equation (2.11). In all, the limit points of the sequence generated by Algorithm 1 1 are stationary points of the gradient dynamical system (4.7) and (4.9).

By the way they are defined, the vector fields on the right hand sides of (4.7) and (4.9), respectively, are analytic in $\mathbf{u}_i^{(1)}, \ldots, \mathbf{u}_i^{(k)}$. As they are gradient flows, by the Łojasiewicz inequality [4, 23] the solution flow $\mathbf{u}_i^{(\ell)}(t)$ must converge to an isolated limit point. That is, the equilibria to the vector field must be isolated. □

4.2. An illustration by matrices. It will be illustrative to observe how the algorithm goes for the case $k = 2$ when T is a matrix. Assuming $s = 2$, the question as we have formulated thus far is to find the best rank-2 approximation while insisting that the second components of the rank-1 matrices are necessarily orthogonal.

Needless to say, the Eckart-Young theorem already asserts that the absolutely best rank-2 approximation should be the truncated singular value decomposition. Applying Algorithm 1, nonetheless, we proceed to update the first components of the rank-1 matrices

$$\begin{cases} \mathbf{u}_{1,[p+1]}^{(1)} & = & \frac{1}{\lambda_{1,[p+1]}^{(1)}} T\mathbf{u}_{1,[p]}^{(2)}, \\ \mathbf{u}_{2,[p+1]}^{(1)} & = & \frac{1}{\lambda_{2,[p+1]}^{(2)}} T\mathbf{u}_{2,[p]}^{(2)}, \end{cases} \tag{4.11}$$

which are followed by the intermediate vectors

$$\begin{cases} \mathbf{v}_{1,[p+1]}^{(2)} & = & T^\top \mathbf{u}_{1,[p+1]}^{(1)}, \\ \mathbf{v}_{2,[p+1]}^{(2)} & = & T^\top \mathbf{u}_{2,[p+1]}^{(1)}. \end{cases} \tag{4.12}$$

Had we stopped here and repeated doing steps (4.11) and (4.12), then this process would have been equivalent to the simultaneous subspace iteration of the power method applied to the matrices $TT^\top$ and $T^\top T$ on the 2-dimensional subspaces $\operatorname{span}\left\{\mathbf{u}_{1,[0]}^{(1)}, \mathbf{u}_{2,[0]}^{(1)}\right\}$ and $\operatorname{span}\left\{\mathbf{u}_{1,[0]}^{(2)}, \mathbf{u}_{2,[0]}^{(2)}\right\}$, respectively, which eventually would have coalesced to the dominant singular vectors of T. Instead, we calculate the parameter t according to (2.17) and define the second components of rank-1 matrices via

$$\begin{cases} \mathbf{u}_{1,[p+1]}^{(2)} & = & \frac{1}{\lambda_{1,[p+1]}^{(2)}} \left(\mathbf{v}_{1,[p+1]}^{(2)} - t\mathbf{v}_{2,[p+1]}^{(2)}\right), \\ \mathbf{u}_{2,[p+1]}^{(2)} & = & \frac{1}{\lambda_{2,[p+1]}^{(2)}} \left(\mathbf{v}_{2,[p+1]}^{(2)} - t\mathbf{v}_{1,[p+1]}^{(2)}\right), \end{cases} \tag{4.13}$$

which orthogonalizes the basis vectors $\left\{\mathbf{v}_{1,[p+1]}^{(2)}, \mathbf{v}_{2,[p+1]}^{(2)}\right\}$ and, hence, prevents the iterates from coalescence.

Note that we do not take any action on orthogonalizing $\mathbf{u}_{1,[p+1]}^{(1)}$ and $\mathbf{u}_{2,[p+1]}^{(1)}$. Indeed, the scheme in Algorithm 1 does not guarantee that they are orthogonal in general. Our optimization subject to the orthogonality of the second components only is taking place over a larger feasible set than that when subject to the complete orthogonality of both components. In general, the more restricted optimization problem should give rise to an inferior objective value. That is not so in the case of rank-2 approximation to matrices. Quite on the contrary, the more restricted orthogonality, that is, the complete orthogonality, produces the singular value decomposition of T. What is truly remarkable about the singular value decomposition of T is that not only it is one of the stationary points to (2.3), but also it is the absolute optimal solution by the Eckart-Young theorem. This absolute optimal approximation automatically entails the complete orthogonality.

For high-order tensors, however, complete orthogonal decomposition does not always exist. So the complete orthogonalization and the absolute optimal approximation are not necessarily consequence of each other for high-order tensors. Considerable endeavors have been devoted to generalizing Eckart-Young theorem to tensors [3, 7, 10, 14, 16, 17, 18, 21, 22, 24] with a variety of applications [25, 28, 26, 27, 30, 32, 34] and the research is still ongoing.

By the ascent property proved in Section 3, we already know that

$$\sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p]}^{(1)}\otimes\mathbf{u}_{i,[p]}^{(2)}\right\rangle^2 \le \sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p+1]}^{(1)}\otimes\mathbf{u}_{i,[p]}^{(2)}\right\rangle^2 \le \sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p+1]}^{(1)}\otimes\mathbf{u}_{i,[p+1]}^{(2)}\right\rangle^2. \tag{4.14}$$

Now we want to give an exact account on how the objective function in (2.3) varies quadratically as the perturbations. The idea is to obtain an estimate of the incremental quantities $\delta\mathbf{u}_{i,[p]}^{(\ell)}$. The first inequality in (4.14) can easily been rewritten as

$$\begin{aligned}\sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p+1]}^{(1)}\otimes\mathbf{u}_{i,[p]}^{(2)}\right\rangle^2 &= \sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p]}^{(1)}\otimes\mathbf{u}_{i,[p]}^{(2)}\right\rangle^2 + \sum_{i=1}^{2}\left\langle T, \delta\mathbf{u}_{i,[p]}^{(1)}\otimes\mathbf{u}_{i,[p]}^{(2)}\right\rangle^2 \\ &\quad +2\sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p]}^{(1)}\otimes\mathbf{u}_{i,[p]}^{(2)}\right\rangle\left\langle T, \delta\mathbf{u}_{i,[p]}^{(1)}\otimes\mathbf{u}_{i,[p]}^{(2)}\right\rangle \\ &= \sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p]}^{(1)}\otimes\mathbf{u}_{i,[p]}^{(2)}\right\rangle^2 + \sum_{i=1}^{2}\left(\lambda_{i,[p+1]}^{(1)}\right)^2\left(1-\frac{\left\|\delta\mathbf{u}_{i,[p]}^{(1)}\right\|^2}{4}\right)\left\|\delta\mathbf{u}_{i,[p]}^{(1)}\right\|^2. \end{aligned}\tag{4.15}$$

In the above, the second equation is obtained by combining the last two terms in the first equation with the substitution $T\mathbf{u}_{i,[p]}^{(2)} = \lambda_{1,[p+1]}\mathbf{u}_{1,[p+1]}^{(1)}$ followed by the fact of Lemma 4.1. Note that $\left\|\delta\mathbf{u}_{i,[p]}^{(1)}\right\| \le 2$, so the term of the second summation (4.15) is nonnegative.

We have already shown the monotone convergence of the sequence $\left\{\sum_{i=1}^{2}\left\langle T, \mathbf{u}_{i,[p]}^{(1)}\otimes\mathbf{u}_{i,[p]}^{(2)}\right\rangle^2\right\}$. From (4.15), it follows that the sequences $\left\{\delta\mathbf{u}_{i,[p]}^{(1)}\right\}$, $i = 1, 2$, must converge to zero as p goes to infinity. This merely fact suffices to set off a chain of events. First, by Lemmas 4.2 and 4.3 (see also [33, Lemmas 3.5 and 3.6]) we conclude that the sequences $\left\{\mathbf{u}_{i,[p]}^{(1)}\right\}$, $i = 1, 2$, also converge. Then, by the relationships (4.12) and (4.13), respectively, we can assert the convergence of the sequences $\left\{\mathbf{v}_{i,[p]}^{(2)}\right\}$ and $\left\{\mathbf{u}_{i,[p]}^{(2)}\right\}$, $i = 1, 2$. Thus the global convergence of iterates by Algorithm 1 applied to a matrix is established.

We mention that proving the convergence of $\left\{\delta\mathbf{u}_{i,[p]}^{(2)}\right\}$ is not to be as directly as (4.15). The twist involved in (4.13) for the purpose of orthogonalization significantly complicates the needed manipulation. However, here is the important lesson we learn which is also applicable to the next section for the general case, namely, the assumption $s = 2$ is immaterial and can easily be circumvented. All we need to do is to take the viewpoint that, in the case that $s = 1$, we simply rearrange the order of multiplications by switching the roles of T and $T^\top$ and regarding the originally first component as the second component.

4.3. General case. The technique developed above for matrices is readily generalizable to tensors of general order k. As we have indicated earlier, by rearranging the order of multiplications if necessary, we may assume without loss of generality that $s = k$. Our goal is to show that all sequences $\left\{\delta\mathbf{u}_{i,[p]}^{(\ell)}\right\}$, $\ell = 1, \ldots k-1$, $i = 1, 2$, converge to zero as p goes to infinity. The following relationship is parallel to that of (4.15).

$$\begin{aligned}
&\sum_{i=1}^{2}\left\langle T, \bigotimes U_{i,[p+1]}^{\{1..\ell)\}} \otimes \bigotimes U_{i,[p]}^{\{(\ell+1)..k\}}\right\rangle^2 = \sum_{i=1}^{2}\left\langle T, \bigotimes U_{i,[p+1]}^{\{1..(\ell-1)\}} \otimes \left(\mathbf{u}_{i,[p]}^{(\ell)} + \delta\mathbf{u}_{i,[p]}^{(\ell)}\right) \otimes \bigotimes U_{i,[p]}^{\{(\ell+1)..k\}}\right\rangle^2 \\
&= \sum_{i=1}^{2}\left\langle T, \bigotimes U_{i,[p+1]}^{\{1..(\ell-1)\}} \otimes \bigotimes U_{i,[p]}^{\{\ell..k\}}\right\rangle^2 + \sum_{i=1}^{2}\left\langle \delta\mathbf{u}_{i,[p]}^{(\ell)}, \underbrace{T \circledast_\ell \bigotimes U_{i,[p+1]}^{\{1..(\ell-1)\}} \otimes \widehat{\mathbf{u}}^{(\ell)} \otimes \bigotimes U_{i,[p]}^{\{(\ell+1)..k\}}}_{\lambda_{i,[p+1]}^{(\ell)}\mathbf{u}_{i,[p+1]}^{(\ell)}}\right\rangle^2 \\
&\quad + 2\sum_{i=1}^{2}\left\langle \mathbf{u}_{i,[p]}^{(\ell)}, \lambda_{i,[p+1]}^{(\ell)}\mathbf{u}_{i,[p+1]}^{(\ell)}\right\rangle\left\langle \delta\mathbf{u}_{i,[p]}^{(\ell)}, \lambda_{i,[p+1]}^{(\ell)}\mathbf{u}_{i,[p+1]}^{(\ell)}\right\rangle \\
&= \sum_{i=1}^{2}\left\langle T, \bigotimes U_{i,[p+1]}^{\{1..(\ell-1)\}} \otimes \bigotimes U_{i,[p]}^{\{\ell..k\}}\right\rangle^2 + \sum_{i=1}^{2}\left(\lambda_{i,[p+1]}^{(\ell)}\right)^2\left(1 - \frac{\left\|\delta\mathbf{u}_{i,[p]}^{(\ell)}\right\|^2}{4}\right)\left\|\delta\mathbf{u}_{i,[p]}^{(\ell)}\right\|^2. \qquad (4.16)
\end{aligned}$$

The fact that $\ell \neq s$ is important because in the third equation above we have used the definition of $\mathbf{u}_{i,[p+1]}^{(\ell)}$ specified in Algorithm 1. Similar to (4.15), the last two equations are then combined by using Lemma 4.1. The first two summations on both sides of (4.16) being sandwiched in between two consecutive elements of the sequence $\left\{\sum_{i=1}^{2}\left\langle T, U_{i,[p]}^{\{1..k\}}\right\rangle^2\right\}$, we must have the following fact.

LEMMA 4.4. *The increment* $\left\{\delta\mathbf{u}_{i,[p]}^{(\ell)}\right\}$ *converges to zero for all* $\ell = 1, \ldots k-1$.

It only remains to justify the convergence of $\left\{\delta\mathbf{u}_{i,[p]}^{(k)}\right\}$. Similar to the discussion in Section 4.2, we first establish the convergence of the sequences $\left\{\mathbf{u}_{i,[p]}^{(\ell)}\right\}$, $\ell = 1, \ldots, k-1$, $i = 1, 2$. But it follows precisely from Lemma 4.2, knowing Lemma 4.4. Then, we employ the relationships (4.12) and (4.13), respectively, to assert the convergence of the sequences $\left\{\mathbf{v}_{i,[p]}^{(k)}\right\}$ and $\left\{\mathbf{u}_{i,[p]}^{(k)}\right\}$, $i = 1, 2$. The convergence for general tensors is now complete.

Two points are worth noting. First, we have showed in the discussion above that the higher-order power method, that is, Algorithm 1, converges globally from any starting points. We stress that this convergence refers to more than that of the objective values. Furthermore, the relationship (4.16) gives an exact characterization on the improvement of the objective value in terms of the incremental change of the variable. Second, the orthogonal rank-2 approximation for tensors is a nonlinear least squares problem. The objective function f defined in (2.1) generally has multiple local minimizers (and thus multiple local maximizers for $\lambda_1^2 + \lambda_2^2$). A limit point obtained by the high-order power method depends on the starting points and is not guaranteed to be the best rank-2 approximation to T [29]. Some starting strategies similar to those proposed in [33] might help to improve this approximation, which should be of an interesting research topic by itself.

5. Conclusion. A higher-order tensor does not necessarily have an optimal low rank approximation. Imposing a minimum requirement on orthogonality among merely two components of the rank-1 tensors changes the situation to an optimization over a compact set and, hence, guarantees the existence of the best orthogonal low rank approximation. In this paper, we propose an alternating least squares algorithm to compute an orthogonal rank-2 approximation. Our contributions are threefold. First, we provide a mechanism to ensure the orthogonality condition while performing the high-order power method. Second, we offer a rigorous proof on the global convergence of this method. Third, we prove that the stationary points are isolated.

REFERENCES

[1] J. CARROLL AND J.-J. CHANG, *Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition*, Psychometrika, 35 (1970), pp. 283–319. 10.1007/BF02310791.

[2] J. CHANG, W. SUN, AND Y. CHEN, *A modified newton's method for best rank-one approximation to tensors*, Applied Mathematics and Computation, 216 (2010), pp. 1859 – 1867. <ce:title>Workshop on Iterative Methods and Preconditioning Techniques</ce:title>.

[3] J. CHEN AND Y. SAAD, *On the tensor SVD and the optimal low rank orthogonal approximation of tensors*, SIAM J. Matrix Anal. Appl., 30 (2008/09), pp. 1709–1734.

[4] R. CHILL, *On the łojasiewicz-Simon gradient inequality*, J. Funct. Anal., 201 (2003), pp. 572–601.

[5] L. X. COMON, P. AND A. L. F. DE ALMEIDA, *Tensor decompositions, alternating least squares and other tales*, J. Chemometrics, 23 (2009), pp. 393–405.

[6] P. COMON, G. GOLUB, L.-H. LIM, AND B. MOURRAIN, *Symmetric tensors and symmetric tensor rank*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1254–1279.

[7] L. DE LATHAUWER, B. DE MOOR, AND J. VANDEWALLE, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278 (electronic).

[8] ———, *On the best rank-1 and rank-$(R_1, R_2, \cdots, R_N)$ approximation of higher-order tensors*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1324–1342 (electronic).

[9] V. DE SILVA AND L.-H. LIM, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1084–1127.

[10] P. DRINEAS AND M. W. MAHONEY, *A randomized algorithm for a tensor-based generalization of the singular value decomposition*, Linear Algebra and its Applications, 420 (2007), pp. 553–571.

[11] N. K. M. FABER, R. BRO, AND P. K. HOPKE, *Recent developments in candecomp/parafac algorithms: a critical review*, Chemometrics and Intelligent Laboratory Systems, 65 (2003), pp. 119 – 137.

[12] R. HARSHMAN, *Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis*, UCLA Working Papers in Phonetics, 16 (1970).

[13] F. HITCHCOCK, *The Expression of a Tensor Or a Polyadic as a Sum of Products*, Contributions from the Department of Mathematics, sn., 1927.

[14] M. ISHTEVA, P.-A. ABSIL, S. VAN HUFFEL, AND L. DE LATHAUWER, *Best low multilinear rank approximation of higher-order tensors, based on the riemannian trust-region scheme*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 115–135.

[15] H. A. L. KIERS, *Towards a standardized notation and terminology in multiway analysis*, Journal of Chemometrics, 14 (2000), pp. 105–122.

[16] E. KOFIDIS AND P. A. REGALIA, *On the best rank-1 approximation of higher-order supersymmetric tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 863–884.

[17] T. G. KOLDA, *Orthogonal tensor decompositions*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 243–255 (electronic).

[18] ———, *A counterexample to the possibility of an extension of the eckart-young low-rank approximation theorem for the orthogonal rank tensor decomposition*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 762–767.

[19] T. G. KOLDA AND B. W. BADER, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

[20] J. B. KRUSKAL, *Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics*, Linear Algebra and its Applications, 18 (1977), pp. 95 – 138.

[21] D. LEIBOVICI AND R. SABATIER, *A singular value decomposition of a k-way array for a principal component analysis of multiway data, pta-k*, Linear Algebra and its Applications, 269 (1998), pp. 307 – 329.

[22] L.-H. LIM, *Singular values and eigenvalues of tensors: a variational approach*, in Computational Advances in Multi-Sensor Adaptive Processing, 2005 1st IEEE International Workshop on, dec. 2005, pp. 129 –132.

[23] S. ŁOJASIEWICZ AND M.-A. ZURRO, *On the gradient inequality*, Bull. Polish Acad. Sci. Math., 47 (1999), pp. 143–145.

[24] C. D. M. MARTIN AND C. F. VAN LOAN, *A jacobi-type method for computing orthogonal tensor decompositions*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1219–1232.

[25] D. MUTI AND S. BOURENNANE, *Multidimensional filtering based on a tensor approach*, Signal Process., 85 (2005), pp. 2338–2353.

[26] P. PAATERO, *The multilinear engine: A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model*, Journal of Computational and Graphical Statistics, 8, pp. pp. 854–888.

[27] B. SAVAS AND L. ELDÉN, *Handwritten digit classification using higher order singular value decomposition*, Pattern Recognition, 40 (2007), pp. 993 – 1003.

[28] A. SHASHUA AND A. LEVIN, *Linear image coding for regression and classification using the tensor-rank principle*, in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1, 2001, pp. I–42 – I–49 vol.1.

[29] A. STEGEMAN AND P. COMON, *Subtracting a best rank-1 approximation may increase tensor rank*, Linear Algebra Appl., 433 (2010), pp. 1276–1300.

[30] J.-T. SUN, H.-J. ZENG, H. LIU, Y. LU, AND Z. CHEN, *Cubesvd: a novel approach to personalized web search*, in Proceedings of the 14th international conference on World Wide Web, WWW '05, New York, NY, USA, 2005, ACM, pp. 382–390.

[31] L. TUCKER, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311. 10.1007/BF02289464.

[32] M. A. O. VASILESCU AND D. TERZOPOULOS, *Multilinear analysis of image ensembles: Tensorfaces*, in Proceedings of the 7th

European Conference on Computer Vision-Part I, ECCV '02, London, UK, UK, 2002, Springer-Verlag, pp. 447–460.
[33] L. WANG AND M. T. CHU, *On the global convergence of the high-order power method for rank-one tensor approximation*, preprint, North Carolina State University, 2013.
[34] M. WELLING AND M. WEBER, *Positive tensor factorization*, Pattern Recogn. Lett., 22 (2001), pp. 1255–1261.
[35] L.-H. ZHANG AND M. T. CHU, *Computing absolute maximum correlation*, IMA J. Numer. Anal., 32 (2012), pp. 163–184.
[36] T. ZHANG AND G. H. GOLUB, *Rank-one approximation to high order tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 534–550.