ON THE TENSOR SVD AND OPTIMAL LOW RANK ORTHOGONAL APPROXIMATIONS OF TENSORS*

JIE CHEN† AND YOUSEF SAAD†

Abstract. It is known that a high order tensor does not necessarily have an optimal low rank approximation, and that a tensor might not be orthogonally decomposable (i.e., admit a tensor SVD). We provide several sufficient conditions which lead to the failure of the tensor SVD, and characterize the existence of the tensor SVD with respect to the Higher Order SVD (HOSVD) of a tensor. In face of these difficulties to generalize standard results known in the matrix case to tensors, we consider low rank orthogonal approximations of tensors. The existence of an optimal approximation is theoretically guaranteed under certain conditions, and this optimal approximation yields a tensor decomposition where the diagonal of the core is maximized. We present an algorithm to compute this approximation and analyze its convergence behavior.

Key words. multilinear algebra, singular value decomposition, tensor decomposition, low rank approximation

AMS subject classifications. 15A69, 15A18

1. Introduction.

There has been renewed interest in studying the properties and decompositions of tensors (also known as N-way arrays or multidimensional arrays) in numerical linear algebra in recent years [25, 24, 38, 31, 8, 17, 18, 26, 22]. The tensor approximation techniques have been fruitfully applied in various areas which include among others, chemometrics [32, 4], signal processing [21, 7], vision and graphics [35, 36], and network analysis [19, 1]. From the point of view of practical applications, the matrix SVD and optimal rank-r approximation of matrices (a.k.a. Eckart–Young theorem [9]) are of particular interest, and it would be nice if these properties could be directly generalized to higher order tensors. However, for any order $N \geq 3$, Silva and Lim [31] showed that the problem of optimal low rank approximation of higher order tensors is ill-posed for many ranks r. Also, Kolda presented numerous examples to illustrate the difficulties of orthogonal tensor decompositions [17, 18]. These studies reveal many aspects of the dissimilarity between tensors and matrices, in spite of the fact that higher order tensors are multidimensional generalizations of matrices.

Currently the most commonly used generalization of the matrix SVD to higher order tensors is the so-called *Higher Order Singular Value Decomposition* (HOSVD) [24]. HOSVD decomposes an order-N tensor into a core tensor that is of the same shape as the original tensor, together with N orthogonal[1] side-matrices. Although this decomposition preserves many nice aspects of the matrix SVD (e.g., the core has the all-orthogonality and the ordering property), a big difference is that the core is in general not diagonal. Hence, in contrast with the matrix SVD, HOSVD cannot be written as a sum of a few orthogonal outer-product terms[2].

*This work was supported by NSF under grants DMS-0510131 and DMS-0528492 and by the Minnesota Supercomputing Institute.

†Department of Computer Science and Engineering, University of Minnesota at Twin Cities, MN 55455. Email: {jchen, saad}@cs.umn.edu.

[1]Throughout this paper, we will say that a matrix $A \in \mathbb{R}^{m\times n}$, with $m \geq n$, is *orthogonal* if $A^T A = I$, in preference to the more common term *unitary*.

[2]Given two sets of vectors $\{u_i\}$ and $\{v_i\}$, the two outer products $u_1 \otimes u_2 \otimes \cdots \otimes u_N$ and $v_1 \otimes v_2 \otimes \cdots \otimes v_N$ are said to be *orthogonal* if $u_i \perp v_i$ for all i. This is also known as *complete orthogonality* in [17]. More on this shortly.

There are three well-known approximations to higher order tensors: (1) rank-1 approximation [25, 38]; (2) rank-$(r_1, r_2, \ldots, r_N)$ approximation with a full core and N orthogonal side-matrices (in the *Tucker/HOOI* fashion) [34, 25]; and (3) approximations using r outer-product terms (in the *CANDECOMP/PARAFAC* fashion) [5, 10]. Among these, only the rank-1 approximation is theoretically guaranteed to have a global optimum [31]. In practice, these approximations are computed using an alternating least squares (ALS) method, where the convergence behavior is theoretically unknown except under a few strong conditions [20]. It has long been observed that the ALS method for the PARAFAC model may converge extremely slowly if at all [28, 15]. An illustration of this phenomenon is given in the Appendix.

In computing the rank-1 approximation, Zhang and Golub [38] presented a generalized Rayleigh quotient iteration that is guaranteed to converge quadratically when it is localized. Alternatively, the approximation using r outer-product terms can be computed in a greedy but suboptimal fashion: An optimal rank-1 approximation is computed and subtracted from the original tensor, yielding the so-called *residual tensor*. Then a rank-1 approximation to the residual tensor is computed, and the residual tensor is updated. This is iterated r times to form the approximation. This approximation can be written in the form of a diagonal core tensor with N side-matrices. However, there is no guarantee that these side-matrices have full column ranks even when r is small. Moreover, the rank of the approximation might be less than r.

Kolda [17] investigated several orthogonal decompositions of tensors related to different definitions of orthogonality, including *orthogonal rank decomposition*, *complete orthogonal rank decomposition* and *strong orthogonal rank decomposition*. These decompositions might not be unique, or even exist. Among these definitions, only the *complete orthogonality* gives a situation which parallels that of the matrix SVD. This approach demands that the side-matrices all be orthogonal. In this paper, we will use the term *tensor singular value decomposition* (tensor SVD, c.f. Definition 5.1) for complete orthogonal rank decomposition. Zhang and Golub [38] proved that for all tensors of order $N \geq 3$, the tensor SVD is unique (up to signs) if it exists, and that the incremental rank-1 approximation procedure will compute this decomposition.

The contribution of this paper is three-fold. First, we give some sufficient conditions indicating which tensors fail to have a tensor SVD. These conditions are related to the rank, the order, and the dimensions of a tensor, hence can be viewed as generalizations of results given in the literature with specific examples. Furthermore, the existence of tensor SVD can be characterized by the diagonality of the core in the HOSVD of the tensor. Second, we discuss a form of low rank approximations—one that requires diagonal core and orthogonal side-matrices. Theoretically the global optimum of this approximation can be attained for any (appropriate) rank. We present an iterative algorithm to compute this approximation. This algorithm can directly be applied to symmetric tensors, whose approximation requires the side-matrices for all modes be the same. Third, the proposed approximation at the maximally possible rank can be equivalently transformed to a decomposition of the tensor, where the diagonal of the core is maximized. This 'maximal diagonality' for symmetric order-3 tensors has been previously investigated in [6] and [23], but our discussion is in a more general context.

2. Tensor Algebra. In this section, we briefly review some concepts and notions that are used throughout the paper. A *tensor* is a multidimensional array of data whose elements are referred by using multiple indices. The number of indices required

is called the *order* of a tensor. We use

$$\mathcal{A} = (a_{i_1,i_2,\ldots,i_N}) \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$$

to denote a tensor $\mathcal{A}$ of order N. For $n = 1, 2, \ldots, N$, d_n is the n-th *dimension* of $\mathcal{A}$. A vector is an order-1 tensor and a matrix is an order-2 tensor. The following illustrates an order-3 tensor, whose first and second indices vary from top to bottom and from left to right (same convention as for a matrix), and the third index varies from front to back:

$$\mathcal{T} = \begin{array}{ccc} a_{112} & \text{---} & a_{122} \\ /\ \ | & & /\ \ | \\ a_{111} & \text{---} & a_{121} \\ |\ \ | & & |\ \ | \\ & a_{212} \text{ —} & | \text{ — } a_{222} \\ |\ / & & |\ / \\ a_{211} & \text{---} & a_{221} \end{array} = \begin{array}{ccc} 5 & \text{---} & 7 \\ /\ \ | & & /\ \ | \\ 1 & \text{---} & 3 \\ |\ \ | & & |\ \ | \\ & 6 \text{ —} & | \text{ — } 8 \\ |\ / & & |\ / \\ 2 & \text{---} & 4 \end{array}.$$

Unfolding and mode-n products. It is hard to visualize tensors of order $N > 3$. They can be flexibly represented when 'unfolded' into matrices. The *unfolding* of a tensor along *mode* n is a matrix of dimension $d_n \times (d_{n+1} \cdots d_N d_1 \cdots d_{n-1})$. We denote the mode-$n$ unfolding of tensor $\mathcal{A}$ by $A_{(n)}$. Each column of $A_{(n)}$ is a column of $\mathcal{A}$ along the n-th mode. For the above example tensor $\mathcal{T}$, the three mode-n unfoldings are

$$T_{(1)} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{bmatrix},$$

$$T_{(2)} = \begin{bmatrix} 1 & 5 & 2 & 6 \\ 3 & 7 & 4 & 8 \end{bmatrix},$$

$$T_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}.$$

An important operation for a tensor is the *tensor-matrix multiplication*, also known as *mode-n product*. Given a tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ and a matrix $M \in \mathbb{R}^{c_n \times d_n}$, the mode-$n$ product is a tensor

$$\mathcal{B} = \mathcal{A} \otimes_n M \in \mathbb{R}^{d_1 \times \cdots \times d_{n-1} \times c_n \times d_{n+1} \cdots \times d_N}$$

where

$$b_{i_1,\ldots,i_{n-1},j_n,i_{n+1},\ldots,i_N} := \sum_{i_n=1}^{d_n} a_{i_1,\ldots,i_{n-1},i_n,i_{n+1},\ldots,i_N} m_{j_n,i_n}$$

for $j_n = 1, 2, \ldots, c_n$. In matrix representation, this is

$$B_{(n)} = M A_{(n)}. \tag{2.1}$$

Two notable properties of tensor-matrix multiplication are:

(i) For $m \neq n$, and matrices F and G of appropriate dimensions,

$$(\mathcal{A} \otimes_n F) \otimes_m G = (\mathcal{A} \otimes_m G) \otimes_n F.$$

(ii) For any n, and for matrices F and G of appropriate dimensions,

$$(\mathcal{A} \otimes_n F) \otimes_n G = \mathcal{A} \otimes_n (GF).$$

Since for a general n, the mode-n product of two matrices is not defined, we can safely omit the parentheses and write $(\mathcal{A} \otimes_n F) \otimes_m G$ as $\mathcal{A} \otimes_n F \otimes_m G$.

Inner products and tensor norm. The *inner product* of two tensors $\mathcal{A}$ and $\mathcal{B}$ of the same shape is defined by

$$\langle \mathcal{A}, \mathcal{B} \rangle_F := \sum_{i_N=1}^{d_N} \cdots \sum_{i_1=1}^{d_1} a_{i_1,\ldots,i_N} b_{i_1,\ldots,i_N}.$$

and the *norm* induced from this inner product is

$$\|\mathcal{A}\|_F := \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle_F}.$$

We say that $\mathcal{A}$ is a *unit tensor* if $\|\mathcal{A}\|_F = 1$. When $N = 2$, $\mathcal{A}$ is a matrix, and $\|\mathcal{A}\|_F$ is its Frobenius norm. The norm of a tensor is equal to the Frobenius norm of the unfolding of the tensor along any mode:

$$\|\mathcal{A}\|_F = \left\|A_{(n)}\right\|_F, \quad \text{for } n = 1, \ldots, N.$$

From the matrix representation of mode-n products (c.f. equation (2.1)), one can easily verify two properties of the tensor norm:
Orthogonal invariance: For any orthogonal matrix $Q \in \mathbb{R}^{c_n \times d_n}$ $(c_n \geq d_n)$,

$$\|\mathcal{A}\|_F = \|\mathcal{A} \otimes_n Q\|_F.$$

Consistency: For any matrix $M \in \mathbb{R}^{c_n \times d_n}$,

$$\|\mathcal{A} \otimes_n M\|_F \leq \|\mathcal{A}\|_F \|M\|_F.$$

Outer product tensors. The outer product of N (column) vectors generalizes standard outer product of two vectors. The outer product of N (column) vectors $x_n \in \mathbb{R}^{d_n}$, is a tensor of dimension $d_1 \times d_2 \times \cdots \times d_N$ which is expressed as

$$\mathcal{X} = x_1 \otimes x_2 \otimes \cdots \otimes x_N,$$

and whose $(i_1, i_2, \ldots, i_N)$-entry is $\prod_{n=1}^{N} (x_n)_{i_n}$, where $(v)_j$ denotes the j-th entry of vector v. It can be verified that the mode-n product of an outer product tensor $\mathcal{X}$ with a matrix M can be computed as follows:

$$\mathcal{X} \otimes_n M = x_1 \otimes \cdots \otimes x_{n-1} \otimes (Mx_n) \otimes x_{n+1} \cdots \otimes x_N,$$

and that the inner product of $\mathcal{X}$ with a general tensor $\mathcal{A}$ is

$$\begin{aligned} \langle \mathcal{A}, \mathcal{X} \rangle_F &= \langle \mathcal{A}, x_1 \otimes x_2 \otimes \cdots \otimes x_N \rangle_F \\ &= \mathcal{A} \otimes_1 x_1^T \otimes_2 x_2^T \otimes \cdots \otimes_N x_N^T. \end{aligned}$$

Let $\mathcal{U} = u_1 \otimes u_2 \otimes \cdots \otimes u_N$ and $\mathcal{V} = v_1 \otimes v_2 \otimes \cdots \otimes v_N$ where $u_n, v_n \in \mathbb{R}^{d_n}$ for $n = 1, 2, \ldots, N$. Then

$$\langle \mathcal{U}, \mathcal{V} \rangle_F = \prod_{n=1}^{N} \langle u_n, v_n \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product of two vectors. A consequence of the above relation is that $\|\mathcal{U}\|_F$ is the product of the 2-norms of the vectors u_n.

Orthogonality of tensors. Two tensors $\mathcal{A}$ and $\mathcal{B}$ of the same shape are *F-orthogonal* (*Frobenius orthogonal*) if their inner product is zero, i.e.,

$$\langle \mathcal{A}, \mathcal{B} \rangle_F = 0.$$

For outer product tensors $\mathcal{U} = u_1 \otimes u_2 \otimes \cdots \otimes u_N$ and $\mathcal{V} = v_1 \otimes v_2 \otimes \cdots \otimes v_N$, the above definition implies that they are F-orthogonal if

$$\prod_{n=1}^{N} \langle u_n, v_n \rangle = 0.$$

This leads to other options for defining orthogonality of two outer products. The paper [17] discussed two cases:

1. Complete orthogonality: $\langle u_n, v_n \rangle = 0$ for all $n = 1, \ldots, N$.
2. Strong orthogonality: For all n, either $\langle u_n, v_n \rangle = 0$ or u_n and v_n are collinear, but there is at least one l such that $\langle u_l, v_l \rangle = 0$.

In this paper we will simply use the term *orthogonal* for two outer products that are completely orthogonal.

Tensor decompositions. A *decomposition* of tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ is of the form

$$\mathcal{A} = \mathcal{B} \otimes_1 S_1 \otimes_2 S_2 \otimes \cdots \otimes_N S_N,$$

where $\mathcal{B} \in \mathbb{R}^{c_1 \times c_2 \times \cdots \times c_N}$ is called the *core tensor*, and $S_n \in \mathbb{R}^{d_n \times c_n}$ for $n = 1, \ldots, N$ are called *side-matrices*. Let $S_n = \left[s_n^{(1)}, s_n^{(2)}, \ldots, s_n^{(c_n)} \right]$ for all n, then the decomposition of $\mathcal{A}$ can equivalently be written as a sum of a series of outer product tensors:

$$\mathcal{A} = \sum_{i_N=1}^{c_N} \cdots \sum_{i_1=1}^{c_1} b_{i_1, i_2, \ldots, i_N} s_1^{(i_1)} \otimes s_2^{(i_2)} \otimes \cdots \otimes s_N^{(i_N)}. \tag{2.2}$$

In particular, if $\mathcal{B}$ is diagonal, i.e., $b_{i_1, i_2, \ldots, i_N} = 0$ except when $i_1 = i_2 = \cdots = i_N$, then

$$\mathcal{A} = \sum_{i=1}^{r} b_{ii\ldots i} s_1^{(i)} \otimes s_2^{(i)} \otimes \cdots \otimes s_N^{(i)} \tag{2.3}$$

where $r = \min\{c_1, \ldots, c_N\}$.

In the tensor analysis literature, the term 'decomposition' is often used when 'approximation' is meant instead. The *Tucker3 decomposition* (more commonly termed *Tucker3 model*) is an approximation in the form of the right-hand side of (2.2), for given dimensions $c_1, c_2, \ldots, c_N$. Usually, it is required that $c_n < \text{rank}_n(\mathcal{A})$ for all n, otherwise the problem is trivial. The HOOI approach computes this approximation with an additional property that all the S_n's are orthogonal matrices. The *PARAFAC decomposition* (more commonly termed *PARAFAC model*) is an approximation in the form of the right-hand side of (2.3), for a pre-specified r. Usually, r is smaller than the smallest dimension of all modes of $\mathcal{A}$, although requiring a larger r is also possible in the ALS algorithm. As will be discussed in the next section, the smallest r that satisfies equality (2.3) is the rank of the tensor $\mathcal{A}$. Other types of approximations have also been proposed. *Non-negative tensor factorization* (NTF), which requires that all the elements on the right-hand side of (2.2) or (2.3) are non-negative, has been studied in [37, 30, 16]. In general, gradient approaches are employed to compute the approximations, similar to the techniques used for non-negative matrix factorization (NMF).

3. Tensor Ranks. The rank of a tensor causes difficulties when attempting to generalize matrix properties to higher order tensors. There are several possible generalizations of the notion of rank. The *n-rank* of a tensor $\mathcal{A}$, for $n = 1, \ldots, N$, denoted by $\text{rank}_n(\mathcal{A})$, is the rank of the unfolding $A_{(n)}$:

$$\text{rank}_n(\mathcal{A}) := \text{rank}(A_{(n)}).$$

The *(outer-product) rank* of $\mathcal{A}$, denoted $\text{rank}(\mathcal{A})$, is defined as

$$\text{rank}(\mathcal{A}) := \min \left\{ r \,\middle|\, \exists\, x_1^{(i)}, \ldots, x_N^{(i)}, i = 1, \ldots, r, \text{ s.t. } \mathcal{A} = \sum_{i=1}^{r} x_1^{(i)} \otimes x_2^{(i)} \otimes \cdots \otimes x_N^{(i)} \right\}.$$

Hence, an outer product tensor has rank one, and the rank of a tensor $\mathcal{A}$ is the minimum number of rank-1 tensors that can sum up to $\mathcal{A}$.

There are a few notable differences between the notion of rank for matrices and for higher order tensors:

1. For $N = 2$, i.e., when $\mathcal{A}$ is a matrix, $\text{rank}_1(\mathcal{A})$ is the row rank, $\text{rank}_2(\mathcal{A})$ is the column rank, and $\text{rank}(\mathcal{A})$ is the outer-product rank, and they are all equal. However, for higher order tensors ($N > 2$), in general, the n-ranks are different for different modes n, and they are different from $\text{rank}(\mathcal{A})$. Furthermore, the rank of a matrix A can not be larger than the smallest dimension of both modes of A, but for tensors this is no longer true, i.e., the rank can be larger than the smallest dimension of the tensor.

2. The matrix SVD yields one possible way of writing a matrix as a sum of outer-product terms, and the number of nonzero singular values is equal to the rank of the matrix. However, a tensor SVD does not always exist (c.f. Section 5), but if it indeed does, it is unique up to signs [27, 38] and the number of singular values is equal to the rank of the tensor.

3. It is well-known that the optimal rank-r approximation of a matrix is simply its truncated SVD. However some tensors may fail to have an optimal rank-r approximation. If such an approximation exists, it is unclear whether it can be written in the form of a diagonal core tensor multiplied by orthogonal side-matrices.

Next are some rank lemmas, which were also given in [31]. These lemmas are helpful in understanding various tensor rank related issues. The first lemma indicates that the rank of a tensor can not be smaller than any of its n-ranks:

LEMMA 3.1. *Let $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ be an order-N tensor. Then*

$$\text{rank}_n(\mathcal{A}) \leq \min\{\text{rank}(\mathcal{A}), d_n\}, \quad \textit{for } n = 1, 2, \ldots, N.$$

The next lemma illustrates a way to construct higher order tensors while preserving the rank. For this we need to define *tensor products* of tensors. The tensor product of an order-N tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ and an order-N' tensor $\mathcal{B} \in \mathbb{R}^{c_1 \times c_2 \times \cdots \times c_{N'}}$ is an order-$(N + N')$ tensor

$$\mathcal{C} = \mathcal{A} \otimes \mathcal{B} \in \mathbb{R}^{d_1 \times \cdots \times d_N \times c_1 \times \cdots \times c_{N'}},$$

where

$$c_{i_1, \ldots, i_N, j_1, \ldots, j_{N'}} := a_{i_1, \ldots, i_N} b_{j_1, \ldots, j_{N'}}.$$

The notation $\otimes$ used here is consistent with that for outer products.

LEMMA 3.2. *Let $\mathcal{A}$ be a tensor and x be a non-zero vector. Then*

$$\operatorname{rank}(\mathcal{A}) = \operatorname{rank}(\mathcal{A} \otimes x).$$

The following lemma indicates that given any dimension $d_1 \times d_2 \times \cdots \times d_N$, we can construct a tensor of arbitrary rank $R \leq \min\{d_1, d_2, \ldots, d_N\}$.

LEMMA 3.3. *For $n = 1, \ldots, N$, let $x_n^{(1)}, \ldots, x_n^{(R)} \in \mathbb{R}^{d_n}$ be linearly independent. Then the tensor*

$$\mathcal{A} = \sum_{i=1}^{R} x_1^{(i)} \otimes x_2^{(i)} \otimes \cdots \otimes x_N^{(i)}$$

has rank R.

The *direct sum* of two tensors $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ and $\mathcal{B} \in \mathbb{R}^{c_1 \times c_2 \times \cdots \times c_N}$ of the same order is defined as the tensor

$$\mathcal{C} = \mathcal{A} \oplus \mathcal{B} \in \mathbb{R}^{(d_1+c_1) \times (d_2+c_2) \times \cdots \times (d_N+c_N)},$$

where

$$c_{i_1,\ldots,i_N} := \begin{cases} a_{i_1,\ldots,i_N} & \text{if } i_n \leq d_n \text{ for all } n = 1, \ldots, N; \\ b_{i_1-d_1,\ldots,i_N-d_N} & \text{if } i_n > d_n \text{ for all } n = 1, \ldots, N; \\ 0 & \text{otherwise.} \end{cases}$$

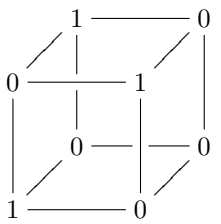
JáJá and Takche [14] showed that if $\mathcal{A}$ and $\mathcal{B}$ are order-3 tensors and at least one of them is a 'stack' of two matrices, then the rank of their direct sum is equal to the sum of their ranks.

THEOREM 3.4 (JáJá–Takche). *Let $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ and $\mathcal{B} \in \mathbb{R}^{c_1 \times c_2 \times c_3}$. If $2 \in \{d_1, d_2, d_3, c_1, c_2, c_3\}$, then*

$$\operatorname{rank}(\mathcal{A} \oplus \mathcal{B}) = \operatorname{rank}(\mathcal{A}) + \operatorname{rank}(\mathcal{B}).$$

4. Ill-Posedness of the Optimal Low Rank Approximation Problem. Silva and Lim [31] proved that for any order $N \geq 3$ and dimensions $d_1, \ldots, d_N \geq 2$, there exists a rank-$(r+1)$ tensor that has no optimal rank-r approximation, for any $r = 2, \ldots, \min\{d_1, \ldots, d_N\}$. This result was further generalized to an arbitrary rank gap, i.e., there exists a rank-$(r+s)$ tensor that has no optimal rank-r approximation, for some r's and s's.

Essentially, this ill-posedness of the optimal approximation problem is illustrated by the fact that the tensor

$\mathcal{E} :=$ 

has rank 3 but can be approximated arbitrarily closely by rank-at-most-2 tensors. Hence $\mathcal{E}$ does not have an optimal rank-2 approximation. Then according to the result of JáJá and Takche and to Lemma 3.2, the ill-posedness of the problem can be generalized to arbitrary rank and order, by constructing higher rank and higher order tensors using direct sums and tensor products. We restate one of the results of [31] in the following theorem. For details of the proof, see the original paper.

THEOREM 4.1. *For $N \geq 3$ and $d_1, d_2, \ldots, d_N \geq 2$, there exists a tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ of rank $r + s$ that has no optimal rank-r approximation, for any r and $s \geq 1$ satisfying $2s \leq r \leq \min\{d_1, d_2, \ldots, d_N\}$.*

5. Tensor SVD and its (Non)Existence. The definition used for the singular value decomposition of a tensor generalizes the matrix SVD from the angle of an expansion of outer product matrices, which becomes an expansion into a sum of high order outer product tensors.

DEFINITION 5.1. *If a tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ can be written in the form*

$$\mathcal{A} = \sum_{i=1}^{R} \sigma_i u_1^{(i)} \otimes u_2^{(i)} \otimes \cdots \otimes u_N^{(i)}, \tag{5.1}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R > 0$ and $\left\langle u_n^{(j)}, u_n^{(k)} \right\rangle = \delta_{jk}$ for $n = 1, 2, \ldots, N$, then (5.1) is said to be the tensor singular value decomposition (tensor SVD) *of $\mathcal{A}$. The σ_i's are* singular values *and the $u_n^{(i)}$'s for $i = 1, \ldots, R$ are the mode-n* singular vectors.

We also call (5.1) the *SVD* of tensor $\mathcal{A}$ for short where there is no ambiguity about tensors and matrices. Expression (5.1) can equivalently be written in the form

$$\mathcal{A} = \mathcal{D} \otimes_1 U_1 \otimes_2 U_2 \otimes \cdots \otimes_N U_N, \tag{5.2}$$

where $\mathcal{D} \in \mathbb{R}^{R \times R \times \cdots \times R}$ is the diagonal *core tensor* with $\mathcal{D}_{ii\ldots i} = \sigma_i$, and

$$U_n = \left[u_n^{(1)}, u_n^{(2)}, \ldots, u_n^{(R)} \right] \in \mathbb{R}^{d_n \times R} \tag{5.3}$$

are orthogonal matrices for $n = 1, 2, \ldots, N$.

Trivially, if a tensor is constructed as in (5.1), its SVD exists. However, in general, a tensor of order $N \geq 3$ may fail to have such a decomposition. In this section, we identify some of these situations.

To begin with, note that the orthogonality requirement of the side-matrices U_n's and Lemma 3.3 imply that the tensor on the right-hand side of (5.1) has rank R. Also, the orthogonality of each U_i implies that $R \leq d_n$ for each n, i.e., $R \leq \min\{d_1, d_2, \ldots, d_N\}$. This leads to the following simple result.

PROPOSITION 5.2. *A tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ with $\operatorname{rank}(\mathcal{A}) > \min\{d_1, d_2, \ldots, d_N\}$ does not admit a tensor SVD.*

Proof. The existence of a tensor SVD such as in (5.1) would trivially lead to a contradiction since the tensor in (5.1) has rank R with $R \leq \min\{d_1, d_2 \cdots, d_N\}$. ∎

Note that Theorem 4.1 guarantees that the condition of Proposition 5.2 is not vacuously satisfied, for any order $N \geq 3$ and dimensions $d_1, d_2, \ldots, d_N \geq 2$.

COROLLARY 5.3. *Given a tensor $\mathcal{A}$ satisfying the condition in Proposition 5.2, any tensor of the form*

$$\mathcal{A} \otimes x_{N+1} \otimes \cdots \otimes x_{N+l},$$

where $l \geq 1$ and $x_{N+1}, \ldots, x_{N+l}$ are nonzero vectors, does not admit a tensor SVD.

Proof. This follows from Proposition 5.2 and Lemma 3.2. ∎

COROLLARY 5.4. *A tensor* $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ *does not admit a tensor SVD if there exists at least one mode* n *such that* $\operatorname{rank}_n(\mathcal{A}) > \min\{d_1, d_2, \ldots, d_N\}$.

Proof. This follows from Proposition 5.2 and Lemma 3.1. ∎

PROPOSITION 5.5. *There exists a tensor* $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ *which does not admit a tensor SVD whenever*

$$d := \max_n\{d_n\} > \min_n\{d_n\} \quad \textit{and} \quad d^2 \leq \prod_{n=1}^{N} d_n.$$

Proof. Without loss of generality, assume that $d = d_1 \geq d_2 \geq \cdots \geq d_N$ and let $d' = d_2 \times \cdots \times d_N$. Since $d \leq d'$, for an arbitrary set of orthonormal vectors $\{a_i \in \mathbb{R}^{d'} \mid i = 1, \ldots, d\}$, we can construct a tensor $\mathcal{A}$ whose unfolding $A_{(1)} = [a_1, a_2, \ldots, a_d]^T$. Then $\operatorname{rank}_1(\mathcal{A}) = d$. By Corollary 5.4, $\mathcal{A}$ does not admit an SVD. ∎

Note that when $N = 2$, i.e., for the matrix case, it is impossible for d_1 and d_2 to satisfy the condition in the proposition.

In closing this section, we provide a necessary and sufficient condition to characterize the existence of the tensor SVD. This is related to the HOSVD proposed by [24]. The essential relation underlying the theorem is that the mode-n singular vectors of $\mathcal{A}$, whose SVD exists, are also the left singular vectors of the unfolding $A_{(n)}$.

THEOREM 5.6. *A tensor* $\mathcal{A}$ *admits an SVD if and only if there exists an HOSVD of* $\mathcal{A}$ *such that the core is diagonal.*

Proof. The sufficient condition is obvious. Consider the necessary condition. If $\mathcal{A}$ can be written in the form (5.1), define the tensor

$$\mathcal{W}_n^{(i)} := u_{n+1}^{(i)} \otimes \cdots \otimes u_N^{(i)} \otimes u_1^{(i)} \otimes \cdots \otimes u_{n-1}^{(i)},$$

and let $w_n^{(i)}$ be the vectorization of $\mathcal{W}_n^{(i)}$. Then the unfolding of $\mathcal{A}$ along mode n is

$$A_{(n)} = \sum_{i=1}^{R} \sigma_i u_n^{(i)} \otimes w_n^{(i)}.$$

Since $\left\langle u_n^{(j)}, u_n^{(k)} \right\rangle = \delta_{jk}$ for all n, we have $\left\langle w_n^{(j)}, w_n^{(k)} \right\rangle = \delta_{jk}$. Hence the above form is the SVD of matrix $A_{(n)}$. In other words, the vectors $u_n^{(1)}, \ldots, u_n^{(R)}$ are the left singular vectors of $A_{(n)}$. From the construction of the HOSVD, expression (5.2) is a valid HOSVD for $\mathcal{A}$. ∎

Due to the non-uniqueness of the matrix SVD, the HOSVD of a tensor may not be unique. Hence even if a tensor is constructed as in (5.1), its HOSVD will not necessarily recover this form.

6. Optimal Low Rank Orthogonal Approximation. The problem addressed by tensor analysis is to approximate some tensor $\mathcal{A}$ by a sum of *simpler* tensors $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_r$. For this it is desirable to minimize

$$\left\| \mathcal{A} - \sum_{i=1}^{r} \sigma_i \mathcal{T}_i \right\|_F$$

for a given r. Without loss of generality, we assume that $\|\mathcal{T}_i\|_F = 1$ for all i. As discussed in Section 4, if the $\mathcal{T}_i$'s are outer product tensors, the infimum of the above

expression might not necessarily be attained. The following proposition reveals some properties when the infimum is indeed achieved.

PROPOSITION 6.1. *Given a tensor $\mathcal{A}$ and an integer r, consider a set of linear combinations of tensors of the form*

$$\mathcal{T} = \sum_{i=1}^{r} \sigma_i \mathcal{T}_i \tag{6.1}$$

where the $\mathcal{T}_i$'s are arbitrary unit tensors. If $\inf \|\mathcal{A} - \mathcal{T}\|_F$ is reached on this set, then for the optimal $\mathcal{T}$ and $\mathcal{T}_i$'s,

$$\langle \mathcal{A} - \mathcal{T}, \mathcal{T}_i \rangle_F = 0 \quad \text{for } i = 1, 2, \ldots, r.$$

Furthermore, if the $\mathcal{T}_i$'s are required to be mutually F-orthogonal, then the optimal σ_i's are related to the optimal $\mathcal{T}_i$'s by

$$\sigma_i = \langle \mathcal{A}, \mathcal{T}_i \rangle_F \quad \text{for } i = 1, 2, \ldots, r. \tag{6.2}$$

In this situation,

$$\|\mathcal{T}\|_F = \sqrt{\textstyle\sum_{i=1}^{r} \sigma_i^2}. \qquad \text{and} \qquad \|\mathcal{A} - \mathcal{T}\|_F^2 = \|\mathcal{A}\|_F^2 - \|\mathcal{T}\|_F^2 . \tag{6.3}$$

Proof. If the infimum is attained by a certain set of σ_i's and $\mathcal{T}_i$'s, and there is a j such that $\langle \mathcal{A} - \mathcal{T}, \mathcal{T}_j \rangle_F = \epsilon \neq 0$, then

$$\begin{aligned}
&\|\mathcal{A} - \textstyle\sum_{i=1}^{r} \sigma_i \mathcal{T}_i - \epsilon \mathcal{T}_j\|_F^2 \\
&= \|\mathcal{A} - \textstyle\sum_{i=1}^{r} \sigma_i \mathcal{T}_i\|_F^2 - 2\epsilon \langle \mathcal{A} - \textstyle\sum_{i=1}^{r} \sigma_i \mathcal{T}_i, \mathcal{T}_j \rangle_F + \epsilon^2 \|\mathcal{T}_j\|_F^2 \\
&= \|\mathcal{A} - \textstyle\sum_{i=1}^{r} \sigma_i \mathcal{T}_i\|_F^2 - \epsilon^2 < \|\mathcal{A} - \textstyle\sum_{i=1}^{r} \sigma_i \mathcal{T}_i\|_F^2 ,
\end{aligned}$$

which contradicts the assumption.

If the unit tensors $\mathcal{T}_i$'s are mutually F-orthogonal, then

$$0 = \langle \mathcal{A} - \textstyle\sum_{i=1}^{r} \sigma_i \mathcal{T}_i, \mathcal{T}_j \rangle_F = \langle \mathcal{A}, \mathcal{T}_j \rangle_F - \sigma_j \langle \mathcal{T}_j, \mathcal{T}_j \rangle_F = \langle \mathcal{A}, \mathcal{T}_j \rangle_F - \sigma_j.$$

and

$$\left\| \mathcal{A} - \sum_{i=1}^{r} \sigma_i \mathcal{T}_i \right\|_F^2 = \|\mathcal{A}\|_F^2 - \sum_{i=1}^{r} 2\sigma_i \langle \mathcal{A}, \mathcal{T}_i \rangle_F + \sum_{i=1}^{r} \sigma_i^2 = \|\mathcal{A}\|_F^2 - \sum_{i=1}^{r} \sigma_i^2 = \|\mathcal{A}\|_F^2 - \|\mathcal{T}\|_F^2 .$$

□

The last part of the proof indicates that the equalities in (6.3) follow from the orthogonality of the $\mathcal{T}_i$'s and the relations (6.2). They do not require optimality.

In this section, we will see that if the $\mathcal{T}_i$'s are orthogonal outer product tensors, then the infimum in the proposition can be attained. Note that in this situation, the approximation $\mathcal{T}$ has rank r. Formally, we will prove that the problem

$$\begin{aligned}
\min \quad & E = \left\| \mathcal{A} - \sum_{i=1}^{r} \sigma_i u_1^{(i)} \otimes u_2^{(i)} \otimes \cdots \otimes u_N^{(i)} \right\|_F \\
\text{s.t.} \quad & \left\langle u_n^{(j)}, u_n^{(k)} \right\rangle = \delta_{jk}, \quad \text{for } n = 1, 2, \ldots, N,
\end{aligned} \tag{6.4}$$

always has a solution for any $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ and any $r \leq \min\{d_1, d_2, \ldots, d_N\}$. The solution for the case $r = \min\{d_1, d_2, \ldots, d_N\}$ leads to a decomposition of $\mathcal{A}$ where the diagonal of the core is maximized.

6.1. Existence of the Global Optimum. Let

$$\mathcal{T}_i := u_1^{(i)} \otimes u_2^{(i)} \otimes \cdots \otimes u_N^{(i)}, \quad \text{for } i = 1, \ldots, r, \tag{6.5}$$

and σ_i's defined as in (6.2), then according to Proposition 6.1 (see comment following the proof),

$$E^2 = \left\| \mathcal{A} - \sum_{i=1}^{r} \sigma_i \mathcal{T}_i \right\|_F^2 = \|\mathcal{A}\|_F^2 - \sum_{i=1}^{r} \sigma_i^2.$$

Hence minimizing E is equivalent to maximizing $\sum_{i=1}^{r} \sigma_i^2$, i.e., the optimization problem (6.4) is equivalent to the following:

$$\begin{aligned} \max \quad & E' = \sum_{i=1}^{r} \left(\mathcal{A} \otimes_1 {u_1^{(i)}}^T \otimes_2 {u_2^{(i)}}^T \otimes \cdots \otimes_N {u_N^{(i)}}^T \right)^2 \\ \text{s.t.} \quad & \left\langle u_n^{(j)}, u_n^{(k)} \right\rangle = \delta_{jk}, \quad \text{for } n = 1, 2, \ldots, N. \end{aligned} \tag{6.6}$$

Let

$$U_n = \left[u_n^{(1)}, u_n^{(2)}, \ldots, u_n^{(r)} \right] \in \Omega_n \tag{6.7}$$

where

$$\Omega_n := \{ W_n \in \mathbb{R}^{d_n \times r} \mid W_n^T W_n = I \} \tag{6.8}$$

for $n = 1, 2, \ldots, N$. The problem (6.6) can be interpreted as that of maximizing E' within the feasible region

$$\Omega := \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N. \tag{6.9}$$

Since for each n the set Ω_n is compact (see, e.g., [13, p. 69]), by Tychonoff Theorem, the feasible region Ω is compact. Under the continuous mapping E', the image $E'(\Omega)$ is also compact. Hence it has a maximum. This proves the following theorem:

THEOREM 6.2. *There exists a solution to the problem* (6.6) *(or equivalently* (6.4) *with σ_i defined in* (6.2)*) for any $r \leq \min\{d_1, d_2, \ldots, d_N\}$.*

6.2. Relation to Tensor Decomposition. Let U_n, $n = 1, \ldots, N$ be the solution to the problem (6.4) with $r = \min\{d_1, d_2, \ldots, d_N\}$ and σ_i be defined in (6.2). Also for $n = 1, \ldots, N$, let $U_n^{\perp}$ be a $d_n \times (d_n - r)$ matrix such that the square matrix

$$\tilde{U}_n := [U_n, U_n^{\perp}] \in \mathbb{R}^{d_n \times d_n} \tag{6.10}$$

is orthogonal. Further, let the tensor

$$\mathcal{S} := \mathcal{A} \otimes_1 \tilde{U}_1^T \otimes_2 \tilde{U}_2^T \otimes \cdots \otimes_N \tilde{U}_N^T \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}. \tag{6.11}$$

Then the equality

$$\mathcal{A} = \mathcal{S} \otimes_1 \tilde{U}_1 \otimes_2 \tilde{U}_2 \otimes \cdots \otimes_N \tilde{U}_N \tag{6.12}$$

holds. This decomposition of $\mathcal{A}$ has the following two properties:

(i) The side-matrices $\tilde{U}_n$ are orthogonal for all n.

(ii) The (squared) norm of the diagonal of the core $\mathcal{S}$:

$$\sum_{i=1}^{\min\{d_1,\ldots,d_N\}} \mathcal{S}_{ii\ldots i}^2 = \sum_{i=1}^{r} \left(\mathcal{A} \otimes_1 u_1^{(i)^T} \otimes_2 u_2^{(i)^T} \otimes \cdots \otimes_N u_N^{(i)^T} \right)^2 = \sum_{i=1}^{r} \sigma_i^2$$

is maximized among all the choices of the orthogonal side-matrices. This is known as *maximal diagonality* in [24].

6.3. First Order Condition. The Lagrangian of (6.6) is

$$L = \sum_{i=1}^{r} \sigma_i^2 - \sum_{j,k=1}^{r} \sum_{n=1}^{N} \mu_n^{j,k} \left(\left\langle u_n^{(j)}, u_n^{(k)} \right\rangle - \delta_{jk} \right), \tag{6.13}$$

where

$$\sigma_i = \mathcal{A} \otimes_1 u_1^{(i)^T} \otimes_2 u_2^{(i)^T} \otimes \cdots \otimes_N u_N^{(i)^T} \tag{6.14}$$

and the $\mu_n^{j,k}$'s are Lagrange multipliers. Define the vector

$$\begin{aligned} v_n^{(i)} &:= \mathcal{A} \otimes_1 u_1^{(i)^T} \otimes \cdots \otimes_{n-1} u_{n-1}^{(i)^T} \otimes_{n+1} u_{n+1}^{(i)^T} \otimes \cdots \otimes_N u_N^{(i)^T} \\ &\in \mathbb{R}^{1\times\cdots\times 1\times d_n\times 1\cdots\times 1}. \end{aligned} \tag{6.15}$$

(Here we abuse the use of notation '='. More precisely, $v_n^{(i)}$ should be the mode-n unfolding of the tensor on the right-hand side of (6.15).) It is not hard to see that $\left\langle u_n^{(i)}, v_n^{(i)} \right\rangle = \sigma_i$ for all n and i, and $v_n^{(i)}$ is the partial derivative of σ_i with respect to $u_n^{(i)}$.

The partial derivative of the Lagrangian with respect to $u_n^{(i)}$ is

$$\frac{\partial L}{\partial u_n^{(i)}} = 2\sigma_i v_n^{(i)} - \sum_{j=1}^{r} \mu_n^{j,i} u_n^{(j)} - \sum_{k=1}^{r} \mu_n^{i,k} u_n^{(k)},$$

for any n and i. By setting the partial derivatives to 0 and putting all equations related to the same n in matrix form, we obtain the following equations:

$$\begin{bmatrix} v_n^{(1)} & \cdots & v_n^{(r)} \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} = \begin{bmatrix} u_n^{(1)} & \cdots & u_n^{(r)} \end{bmatrix} \begin{bmatrix} \frac{\mu_n^{1,1}+\mu_n^{1,1}}{2} & \cdots & \frac{\mu_n^{1,r}+\mu_n^{r,1}}{2} \\ \vdots & \ddots & \vdots \\ \frac{\mu_n^{r,1}+\mu_n^{1,r}}{2} & \cdots & \frac{\mu_n^{r,r}+\mu_n^{r,r}}{2} \end{bmatrix}, \tag{6.16}$$

for all $n = 1, 2, \ldots, N$. Let

$$V_n := \left[v_n^{(1)}, v_n^{(2)}, \ldots, v_n^{(r)} \right], \tag{6.17}$$

$$\Sigma := \operatorname{diag}(\sigma_1, \ldots, \sigma_r), \tag{6.18}$$

and let M_n be the second term on the right-hand side of (6.16). Then (6.16) is compactly represented as

$$V_n \Sigma = U_n M_n, \quad n = 1, 2, \ldots, N. \tag{6.19}$$

In summary, the necessary condition of an extremum of the Lagrangian is the equation (6.19), where V_n is defined in (6.17), Σ is defined in (6.18), U_n is defined in (6.7), and M_n is symmetric, for all $n = 1, 2, \ldots, N$.

We do not consider the Hessian (second order condition), since the feasible region Ω is a subset of the bounded sphere $S^{(d_1+\cdots+d_N)\times r-1}$ with radius $\sqrt{Nr}$, where there does not exist any feasible search direction for any point.

6.4. Algorithm. We seek orthogonal matrices U_n's and symmetric matrices M_n's which satisfy the system (6.19). (The Σ and V_n matrices are computed from the U_n's.) Note that the pair U_n, M_n happens to be the polar decomposition of the matrix $V_n\Sigma$. Hence the system can be solved in an iterative fashion: We begin with an initial guess of the set of orthogonal matrices $\{U_1, U_2, \ldots, U_N\}$, which can be obtained, say, by HOSVD. For each n, we compute V_n and Σ, and update U_n as an orthogonal polar factor of $V_n\Sigma$. This procedure is iterated until convergence is observed. Algorithm 1 (LROAT) summarizes this idea.

Algorithm 1 Low Rank Orthogonal Approximation of Tensors (LROAT)

Input: Tensor $\mathcal{A}$, rank r, orthogonal matrices $U_1, \ldots, U_N$ as initial guess
Output: $\sigma_1, \ldots, \sigma_r$, $U_1, \ldots, U_N$

repeat
  for $n \leftarrow 1, \ldots, N$ **do**
    Compute $V_n = \left[v_n^{(1)}, \ldots, v_n^{(r)}\right]$ according to (6.15)
    Compute $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ according to (6.14)
    $[Q_n, H_n] \leftarrow$ polar-decomp$(V_n\Sigma)$
    Update $U_n \leftarrow Q_n$
  end for
until convergence

6.5. Convergence. Algorithm 1 employs an alternating procedure (iterating through U_1, U_2, ..., U_N), where in each step all but one (U_n) parameters are fixed. In general, algorithms of this type, including alternating least squares, are not guaranteed to converge. Specifically, the objective function may converge but not the parameters. (See, for example, [20] for some discussion.) For Algorithm 1, we are also unable yet to prove global convergence, though empirically it appears to hold. However, in the sequel, we will prove that: (1) The iterations monotonically increase the objective value E' (Theorem 6.4); (2) Under a mild condition, of the generated parameter sequence, every converging subsequence converges to a stationary point of the objective function (Theorem 6.7); and (3) In a neighborhood of a local maximum, the parameter sequence converges to this stationary point (Theorem 6.9).

Before analyzing the convergence behavior of Algorithm 1, we index all the iterates. The outer-loop is indexed by p and the overall iterations are indexed by idx, which is equal to $n + (p-1)N$. In other words, the above algorithm is rewritten as follows.

for $p \leftarrow 1, 2, \ldots$ **do**
  for $n \leftarrow 1, \ldots, N$ **do**
    $idx = n + (p-1)N$
    For all i, compute $\sigma_i^{(idx)}$ according to $U_1^{(p+1)}, \ldots, U_{n-1}^{(p+1)}, U_n^{(p)}, U_{n+1}^{(p)}, \ldots, U_N^{(p)}$

Objective $E'^{(idx)} = \sum_{i=1}^{r} \left(\sigma_i^{(idx)}\right)^2$
Compute $V_n^{(p)}$ from $U_1^{(p+1)}, \ldots, U_{n-1}^{(p+1)}, U_{n+1}^{(p)}, \ldots, U_N^{(p)}$
$\Sigma^{(idx)} = \text{diag}\left(\sigma_1^{(idx)}, \ldots, \sigma_r^{(idx)}\right)$
Polar decomposition $V_n^{(p)}\Sigma^{(idx)} = Q_n^{(p)} H_n^{(p)}$
Update $U_n^{(p+1)} = Q_n^{(p)}$
end for
end for

The following lemma, which is well-known when the matrix A is square, reveals the trace maximizing property that is important for the convergence analysis of Algorithm 1.

LEMMA 6.3. *Let matrix $A \in \mathbb{R}^{m\times n}$, $m \geq n$, have polar decomposition $A = QH$ where $Q \in \mathbb{R}^{m\times n}$ is the orthogonal polar factor and $H \in \mathbb{R}^{n\times n}$ is the symmetric positive semi-definite factor, then*

$$\max_{P\in\mathbb{R}^{m\times n},\, P^TP=I} \text{tr}(P^TA)$$

is attained when $P = Q$.

Proof. Any P can be written as ZQ, where $Z \in \mathbb{R}^{m\times m}$ is orthogonal. Then

$$\text{tr}(P^TA) = \text{tr}(Q^TZ^TQH) = \text{tr}(Z^TQHQ^T).$$

Since QHQ^T is symmetric positive semi-definite, $\max \text{tr}(Z^TQHQ^T)$ is attained when $Z = I$. □

Since $U_n^{(p+1)}$ is the orthogonal polar factor of $V_n^{(p)}\Sigma^{(idx)}$, by Lemma 6.3,

$$\sum_{i=1}^{r} \left(\sigma_i^{(idx)}\right)^2 = \text{tr}\left({U_n^{(p)}}^T V_n^{(p)}\Sigma^{(idx)}\right) \leq \text{tr}\left({U_n^{(p+1)}}^T V_n^{(p)}\Sigma^{(idx)}\right) = \sum_{i=1}^{r} \sigma_i^{(idx+1)}\sigma_i^{(idx)}.$$

Then by the Cauchy-Schwarz inequality,

$$\sum_{i=1}^{r} \left(\sigma_i^{(idx)}\right)^2 \leq \sum_{i=1}^{r} \sigma_i^{(idx+1)}\sigma_i^{(idx)} \leq \sum_{i=1}^{r} \left(\sigma_i^{(idx+1)}\right)^2, \tag{6.20}$$

and

$$\sum_{i=1}^{r} \left(\sigma_i^{(idx)}\right)^2 = \sum_{i=1}^{r} \left(\sigma_i^{(idx+1)}\right)^2 \quad \text{iff} \quad \sigma_i^{(idx)} = \sigma_i^{(idx+1)} \text{ for all } i. \tag{6.21}$$

Inequality (6.20) means that each update of U_n increases the value of the objective function E', i.e.,

$$E'^{(idx)} \leq E'^{(idx+1)}.$$

Since E' is bounded from above (existence of the maximum, c.f. Theorem 6.2), the sequence $\{E'^{(idx)}\}_{idx=1}^{\infty}$ converges. Note that the convergence does not depend on the initial guess input to the algorithm. Formally, we have established the following theorem:

THEOREM 6.4. *Given any initial guess, the iterations of Algorithm 1 monotonically increase the objective function E' defined in* (6.6) *to a limit.*

The convergence of the objective function does not necessarily imply that the function parameters will converge. However, in our case since the parameters U_n's are bounded, they admit converging subsequences. Next we will show that every such subsequence converges to a stationary point of E'. For this, the following lemma uses a helper function f.

LEMMA 6.5. *Let $T : \Theta \to \Theta$ be a continuous mapping and a sequence $\{\theta_n \in \Theta\}_{n=1}^{\infty}$ be generated from the fixed point iteration $\theta_{n+1} = T(\theta_n)$. If there exists a continuous function $f : \Theta \to \mathbb{R}$ satisfying the following two conditions:*

(i) *The sequence $\{f(\theta_n)\}_{n=1}^{\infty}$ converges;*

(ii) *For $\theta \in \Theta$, if $f(T(\theta)) = f(\theta)$ then $T(\theta) = \theta$;*

then every converging subsequence of $\{\theta_n\}_{n=1}^{\infty}$ converges to a fixed point of T.

Proof. Let $\{\theta_{s_n}\}_{n=1}^{\infty}$ be a converging subsequence of $\{\theta_n\}_{n=1}^{\infty}$, where $\theta_{s_n} \to \theta^*$. Also let f^* be the limit of $f(\theta_n)$. Then $f(\theta_{s_n}) \to f(\theta^*)$, therefore $f(\theta^*) = f^*$. Meanwhile from the continuity of T and f, we have $T(\theta_{s_n}) \to T(\theta^*)$ and $f(\theta_{s_n+1}) = f(T(\theta_{s_n})) \to f(T(\theta^*))$, which implies that $f(T(\theta^*)) = f^*$. Condition (ii) of the lemma now implies that $\theta^* = T(\theta^*)$. □

Our objective function E' is just one such helper function f, and the orthogonal polar factor function plays the role of the mapping T in the above lemma. The following lemma establishes the fact that the orthogonal polar factor function is continuous.

LEMMA 6.6. *The orthogonal polar factor function $g : A \to Q$ defined on the set of matrices with full column rank is continuous. Here Q is the orthogonal polar factor of $A \in \mathbb{R}^{m\times n}$, $m \geq n$.*

Proof. First, function g is well defined, since the orthogonal polar factor of a full rank matrix exists and is unique [12]. If Q and Q' are the orthogonal polar factors of A and A', respectively, Sun and Chen [33] have shown that

$$\|Q - Q'\|_F \leq \frac{2}{\|A^+\|_2} \|A - A'\|_F ,$$

where $^+$ means pseudo inverse. Hence if $A_1, A_2, \ldots$ converges to A^*, then $g(A_1), g(A_2), \ldots$ converges to $g(A^*)$. □

Now we are ready to prove the following result.

THEOREM 6.7. *Every converging subsequence of $\{U_1^{(p)}, \ldots, U_N^{(p)}\}_{p=1}^{\infty}$ generated by Algorithm 1 converges to a stationary point of the objective function E' defined in* (6.6)*, provided the matrices V_n in line 3 of the algorithm do not become rank-deficient throughout the iterations.*

Proof. For convenience, let U denote the side-by-side concatenation of the U_n matrices, i.e., at iteration number p we write $U^{(p)} = \left[U_1^{(p)}, \ldots, U_N^{(p)}\right]$. For each iteration, $V_n^{(p)}\Sigma^{(idx)}$ is computed from $U^{(p)}$ and polar factorized, and $U_n^{(p)}$ is updated. Let T be the composite of all these iterations running n from 1 to N. That is, $U^{(p+1)} = T(U^{(p)})$. It is not hard to see that T is continuous by Lemma 6.6. The objective function E' taking parameter $U^{(p)}$ has been previously shown such that the sequence $\{E'(U^{(p)})\}_{p=1}^{\infty}$ is monotonically converging.

Hence by Lemma 6.5, in order to prove this theorem, it will suffice to show that $E'(T(U)) = E'(U)$ implies $T(U) = U$. Then every converging subsequence of $\{E'(U^{(p)})\}_{p=1}^{\infty}$ converges to a fixed point, which satisfies the first order condition (6.19), i.e., it is also a stationary point of E'.

If $E'(T(U)) = E'(U)$, formula (6.21) indicates that the σ_i values have not changed after the iteration. In particular, for any n, the update of U_n has not changed

$\operatorname{tr}\left(U_n^T V_n \Sigma\right)$. Since orthogonal polar factor of $V_n\Sigma$ is unique when V_n is not rank-deficient, this means that U_n has not changed. This in turn means that U is a fixed point of the mapping T. □

The condition in the theorem is not a strong requirement in general. Of course, the columns $v_n^{(i)}$ of the matrix V_n, as computed from (6.15), will be linearly dependent if the n-rank of $\mathcal{A}$ is less than r. For practical applications, the tensor usually has full n-ranks for all n, so this does hamper the applicability of the theorem.

Though the global convergence of the U_n matrices is not determined, when localized, it is possible that this parameter sequence converges. The following lemma and theorem consider this situation.

LEMMA 6.8. *If a sequence $\{\theta_n\}_{n=1}^{\infty}$ is bounded, and all of its converging subsequences converge to θ^*, then $\theta_n \to \theta^*$.*

Proof. (By contradiction.) If $\{\theta_n\}_{n=1}^{\infty}$ does not converge to θ^*, then there is an $\epsilon > 0$ such that there exists a subsequence $S = \{\theta_{s_n}\}_{n=1}^{\infty}$, where $\|\theta_{s_n} - \theta^*\| \geq \epsilon$ for all n. Since S is bounded, it has a converging subsequence S'. Then S' as a subsequence of $\{\theta_n\}_{n=1}^{\infty}$ converges to a limit other than θ^*. □

THEOREM 6.9. *Let $U^* = [U_1^*, \ldots, U_N^*]$ be a local maximum of the objective function E' defined in* (6.6)*. If the sequence $\{U^{(p)} = [U_1^{(p)}, \ldots, U_N^{(p)}]\}_{p=1}^{\infty}$ generated by Algorithm 1 lies in a neighborhood of U^*, where U^* is the only stationary point in that neighborhood, and if the full rank requirement in Theorem 6.7 is satisfied, then the sequence $\{U^{(p)}\}_{p=1}^{\infty}$ converges to U^*.*

Proof. This immediately follows from Theorem 6.7 and Lemma 6.8. □

Note that since the starting elements of a sequence have no effect on its convergence behavior, the above theorem holds whenever the tailing subsequence, starting from a sufficiently large p, lies within the neighborhood.

A weaker, but simpler, result is the following corollary.

COROLLARY 6.10. *Let $U^* = [U_1^*, \ldots, U_N^*]$ be a local maximum of the objective function E' defined in* (6.6)*. If this local maximum is unique and if the full rank requirement in Theorem 6.7 is satisfied, then the sequence $\{U^{(p)}\}_{p=1}^{\infty}$ converges to U^*.*

6.6. Symmetric Tensors. An order-N tensor $\mathcal{A} \in \mathbb{R}^{d \times d \times \cdots \times d}$, whose dimensions of all modes are the same, is *symmetric* if for all permutations π,

$$a_{i_1, i_2, \ldots, i_N} = a_{i_{\pi(1)}, i_{\pi(2)}, \ldots, i_{\pi(N)}}.$$

For symmetric tensors, usually the approximation problem (6.4) has an additional constraint that the side-matrices U_n's are the same for all n, i.e.,

$$\begin{aligned} \min \quad & E = \left\| \mathcal{A} - \sum_{i=1}^{r} \sigma_i u^{(i)} \otimes u^{(i)} \otimes \cdots \otimes u^{(i)} \right\|_F \\ \text{s.t.} \quad & \left\langle u^{(j)}, u^{(k)} \right\rangle = \delta_{jk}. \end{aligned} \tag{6.22}$$

Applying similar arguments to those in Section 6.1, it is easily seen that (6.22) is equivalent to the following problem:

$$\begin{aligned} \max \quad & E' = \sum_{i=1}^{r} \left(\mathcal{A} \otimes_1 {u^{(i)}}^T \otimes_2 {u^{(i)}}^T \otimes \cdots \otimes_N {u^{(i)}}^T \right)^2 \\ \text{s.t.} \quad & \left\langle u^{(j)}, u^{(k)} \right\rangle = \delta_{jk}. \end{aligned} \tag{6.23}$$

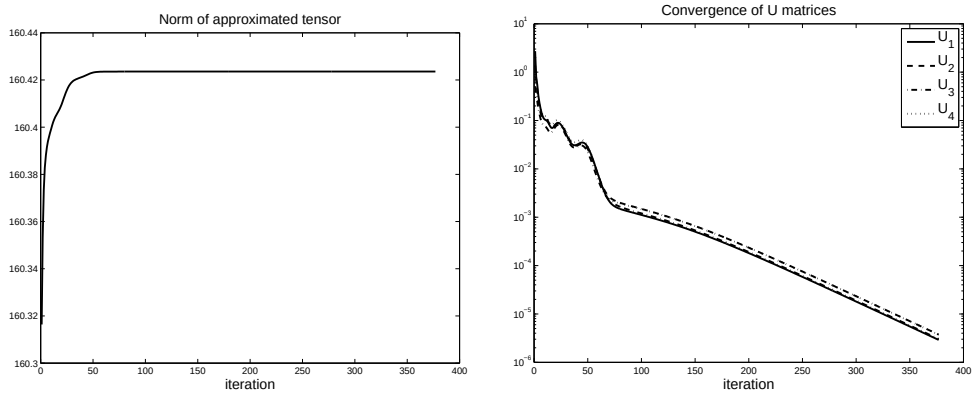
The supremum of E' can be attained. Further, the 'maximal-diagonality' decomposition of $\mathcal{A}$ (c.f. equation (6.12)) has an additional property that the core $\mathcal{S}$ is symmetric. Also, the first order condition (6.19) is simplified to

$$V\Sigma = UM.$$

Algorithm 1 can be directly applied to compute the approximation, except that only a single initial guess U is needed and the for-loop on n (line 2) is omitted. The convergence analysis (Theorems 6.4, 6.7 and 6.9) in the previous subsection also holds.

7. Numerical Experiments. This section will show a few experiments to illustrate the convergence behavior of LROAT (Algorithm 1). The performance of LROAT is compared with that of Tucker and of PARAFAC. For the latter two approximation models we use codes (with modifications) obtained from the MATLAB Tensor Toolbox developed by Bader and Kolda [2]. We use the major left singular vectors of unfoldings of the tensor as the initial guess input for all the algorithms compared. When it comes to the quality of the final approximation, experience shows that compared with random orthonormal vectors, singular vectors as initial guesses do not offer any advantage. It has been argued that running the algorithms several times using different sets of random initial guess enhances the probability of hitting the global optimum. We use singular vectors here only for repeatability.

In the first experiment, we randomly generate a tensor $\mathcal{A}$ of dimension $20 \times 16 \times 10 \times 32$, and use $r = 5$. Figure 7.1(a) shows the norm of the approximated tensor for each iteration. It can be seen that the norm of the approximation increases. Indeed, the norm of initial guess of the tensor is already close to that of the final result. Figure 7.1(b) shows the convergence behavior of the U_n matrices. Since the optima are unknown, for each n, we plot the differences of $\left\| U_n^{(p)} - U_n^{(p-1)} \right\|_F$ for each iteration p. The general shape of the curves seem to indicate a linear convergence for the sequence of matrices U_n.



(a) Norms of the approximated tensors.

(b) Differences of the U_n matrices between consecutive iterations.

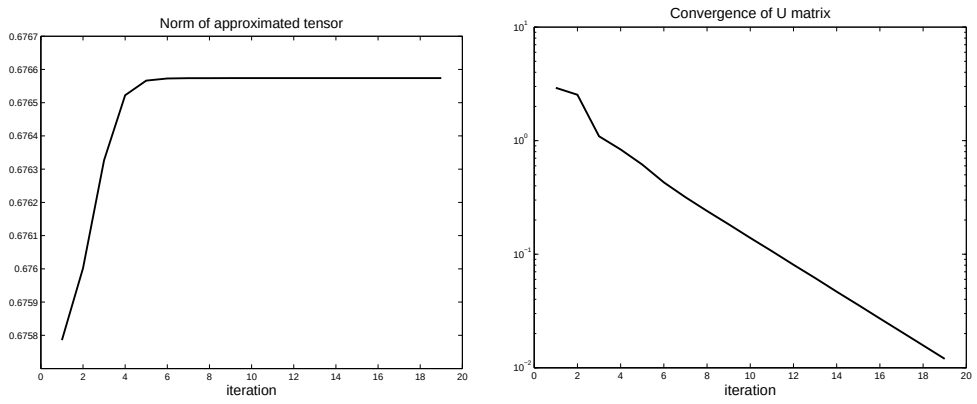
FIG. 7.1. *Experiment 1: Convergence of LROAT for a randomly generated order-4 tensor* $\mathcal{A} \in \mathbb{R}^{20\times16\times10\times32}$.

In the second experiment, we compute the low rank orthogonal approximation of

a symmetric tensor $\mathcal{A}$ which is defined by

$$a_{ijk} = \frac{1}{i^2 + j^2 + k^2}.$$

Remember that for symmetric tensors, there is an additional requirement that all the side-matrices are the same. See Figure 7.2. This figure looks very similar to the one in the previous experiment, and the order of convergence of the U matrix also seems linear. However, by examining the slope of the curve in (b), one sees that the convergence for symmetric tensor approximation is much faster than that of the previous random tensor.



(a) Norms of the approximated tensors.

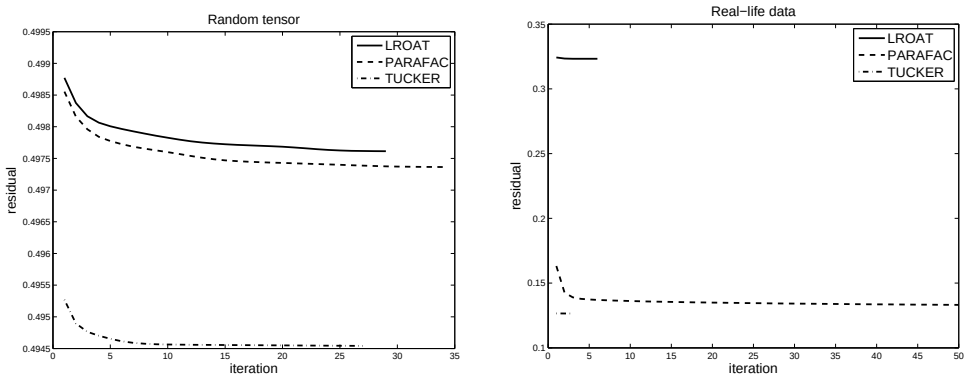
(b) Differences of the U matrix between consecutive iterations.

FIG. 7.2. *Experiment 2: Convergence of LROAT for a symmetric order-3 tensor $\mathcal{A}$ where $a_{ijk} = 1/(i^2 + j^2 + k^2)$.*

In the third experiment, we compare the convergence and approximation quality of three different models: LROAT, Tucker and PARAFAC. See Figure 7.3. Subfigure 7.3(a) shows the typical behavior of a random tensor while 7.3(b) is for a real-life tensor. The latter one is obtained from a problem in acoustics [11], and the data can be downloaded from [3]. The residual norms (fits)

$$fit(p) = \frac{\left\|\mathcal{A} - \mathcal{T}^{(p)}\right\|_F}{\|\mathcal{A}\|_F}$$

over all the iterations p are plotted. From the figure, we see that the residual-norm curves all monotonically decrease, and the steepest descent appears at the beginning few iterations. Theoretically, the optimal residual norm for LROAT should be larger than those of Tucker and PARAFAC. LROAT can be considered a special case of Tucker where the core is full. LROAT is also a special case of PARAFAC where the side-matrices are not restricted to be orthogonal. Hence it is not surprising to see that the curve of LROAT is above those of Tucker and PARAFAC. Two more facts to note are that none of these three models may yield good representation of the original data (in (a), more than 49% of the information is lost), and PARAFAC is usually slow to converge.



(a) Random tensor. $\mathcal{A} \in \mathbb{R}^{10\times30\times20\times15}$. $r = 5$. (b) Real-life tensor. $\mathcal{A} \in \mathbb{R}^{5\times10\times13}$. $r = 3$.

FIG. 7.3. *Experiment 3: Comparison of LROAT, Tucker and PARAFAC.*

8. Concluding Remarks. In the present paper we studied the tensor SVD, and characterized its existence in relation to HOSVD. Similar to the concept of rank, the SVD of high order tensors, exhibits a quite different behavior and characteristics from those of matrices. Thus, the SVD of a matrix is guaranteed to exist, though it may have different representations due to orthogonal transformation of singular vectors corresponding to the same singular value. On the other hand, there are many ways in which a tensor can fail to have an SVD (see the results in Section 5), but when it exists, this decomposition is unique up to signs.

We have also discussed a new form of optimal low rank approximation of tensors, where orthogonality is required. This approximation is inspired by the constraints of Tucker and PARAFAC, and by the ill-posedness of the problem of a general low rank approximation. In some applications, the proposed approximation model may be favored, since it results in N sets of orthonormal vectors or, equivalently, r F-orthogonal unit outer product tensors with different weights. Among the advantages of this approximation over the Tucker model is the fact that it requires far fewer entries to represent the core, and that it is easier to interpret. Also, compared with the PARAFAC model, the orthogonality of vectors may be useful in some cases. Further, LROAT does not seem to exhibit the well-known slow convergence from which PARAFAC suffers.

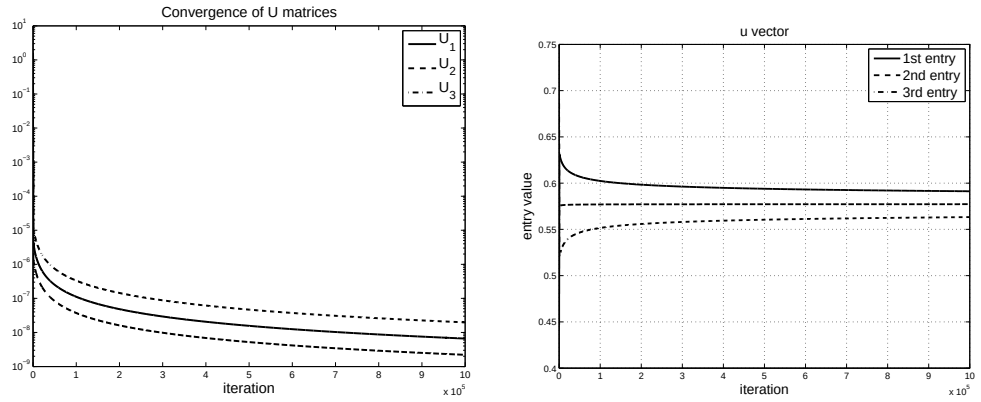
A major restriction of the proposed model is that the number of terms r can not exceed the smallest dimension of all modes of the tensor. A consequence is that the approximation may still be very different from the original tensor even when the maximum r is employed. However we note that when performing data analysis, the interpretation of the vectors and the core tensor is more important than merely focusing on how much is lost when the data is approximated.

A nice aspect of the proposed approximation is that theoretically the optimum of the objective function can be achieved, in contrast to the PARAFAC model which is ill-posed in a strict mathematical sense. We presented an algorithm to compute this approximation, but similar to the algorithms for Tucker and PARAFAC, the computed result is only optimal in a local neighborhood. It will be interesting to study for what initial guesses LROAT converges to the global optimum, or to devise a new algorithm to solve the optimization problem. It is an open problem how fast LROAT converges, although empirically convergence is observed to be linear.

Appendix. Does PARAFAC Converge?. It has been pointed out that the ALS algorithm for computing the PARAFAC model may converge very slowly due to degenerate solutions or multicollinearities, and many alternatives have been proposed to address this problem [28, 29, 15]. During iterations, the objective value monotonically decreases by the nature of the alternating least squares procedure, and since the sequence is bounded, it converges. However, there lacks a rigorous proof about the convergence of the parallel factors. In general it is assumed that these factors converge, but may take a very large number of iterations. In this section, we discuss an experiment showing that the general concept of convergence is unclear in this context. Though only one example is given, we note that the exhibited behavior is not rare for randomly generated tensors. On the other hand it may be argued that tensors in real applications are far from being filled with random numbers.

We generate an order-3 tensor $\mathcal{A} \in \mathbb{R}^{3\times3\times3}$ and run the ALS algorithm on $r = 2$, using $e_1 \otimes e_1 \otimes e_1 + e_2 \otimes e_2 \otimes e_2$ as initial guess. The Matlab code which generates the tensor $\mathcal{A}$ is as follows:

```
A(:,:,1) = [1 2 3; 4 5 6; 7 8 9];
A(:,:,2) = [10 11 12; 13 14 15; 16 17 18];
A(:,:,3) = [19 20 21; 22 23 24; 25 26 27];
```



(a) Differences of the U_n matrices between consecutive iterations.

(b) The entry values of the parallel factor $u_1^{(2)}$ for all iterations.

FIG. A.1. *Slow convergence or non-convergence of PARAFAC.*

Two figures are plotted after running $1,000,000$ iterations. These are shown in Figure A.1. For each of the side-matrices U_1, U_2 and U_3, subfigure A.1(a) shows the norm of the differences between the matrices of two consecutive iterations. It can be seen that the three curves decrease. A necessary condition for PARAFAC to converge is that all these curves decrease to zero. If a curve tends to some nonzero value, this implies that PARAFAC does not converge. We use the following expression

$$\log_{10} y = \frac{a}{(10^{-5}x)^{1/\alpha}} + b$$

to fit the tail curve for U_1 starting at the 2×10^5th iteration. Table A.1 gives some results. When the number of iterations tends to infinity, the value 10^b will show the limit of the difference between two consecutive U_1 matrices.

It is difficult to conclude from this experiment that PARAFAC does not converge for this example since rounding has not been taken into account. However, it makes

TABLE A.1
Curve fitting for different α values.

α	1	2	3	4	5
a	2.3492	2.3098	2.7013	3.1731	3.6732
b	−8.3301	−8.8669	−9.4044	−9.9420	−10.4795
fit error ($\times 10^{-4}$)	0.5550	0.2826	0.1902	0.1439	0.1162

no practical difference for this case whether the sequence actually converges or if it is exceedingly slow to converge. The result will be an inordinate number of iterations to reach a desirable level of convergence, and the cost will not be too high in practice. This can be made evident by examining subfigure A.1(b), which plots the parallel factor $u_1^{(2)}$ over all the iterations. It takes $500,000$ steps for the first entry of $u_1^{(2)}$ to decrease from 0.5939 to 0.5912.

REFERENCES

[1] E. ACAR, S. A. CAMTEPE, M. KRISHNAMOORTHY, AND B. YENER, *Modeling and multiway analysis of chatroom tensors*, in Proc. of IEEE Int. Conf. on Intelligence and Security Informatics (ISI 05), 2005.

[2] B. W. BADER AND T. G. KOLDA, *Algorithm 862: Matlab tensor classes for fast algorithm prototyping*, ACM Trans. Math. Softw., 32 (2006).

[3] D. M. B.L.R., *Daisy: Database for the identification of systems.* `http://homes.esat.kuleuven.be/~smc/daisy/`. Used dataset: `tongue.dat`, section: Biomedical Systems, code: 97-001.

[4] R. BRO, *Review on multiway analysis in chemistry—2000-2005*, Critical Reviews in Analytical Chemistry, 36 (2006), pp. 279–293.

[5] J. D. 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.

[6] P. COMON, *Independent component analysis, a new concept?*, Signal Processing, 36 (1994), pp. 287–314.

[7] ———, *Tensor decompositions*, in Mathematics of Signal Processing V, J. G. McWhirter and I. K. Proudler, eds., Oxford University Press, 2002.

[8] P. COMON, G. GOLUB, L.-H. LIM, AND B. MOURRAIN, *Symmetric tensor and symmetric tensor rank*, SIAM J. Matrix Anal. Appl., (to appear).

[9] C. ECKART AND G. YOUNG, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[10] R. HARSHMAN, *Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis*, UCLA Working Papers in Phonetics, 16 (1970), pp. 1–84.

[11] R. HARSHMAN, P. LADEFOGED, AND L. GOLDSTEIN, *Factor analysis of tongue shapes*, J. Acoust. Soc. Am., 62 (1977), pp. 693–707.

[12] N. J. HIGHAM AND R. S. SCHREIBER, *Fast polar decomposition of an arbitrary matrix*, SIAM J. Sci. Comput., 11 (1990), pp. 648–655.

[13] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, 1985.

[14] J. JÁJÁ AND J. TAKCHE, *On the validity of the direct sum conjecture*, SIAM J. Comput., 15 (1986), pp. 1004–1020.

[15] H. A. L. KIERS, *A three-step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity*, J. Chemometrics, 12 (1998), pp. 155–171.

[16] Y.-D. KIM AND S. CHOI, *Nonnegative tucker decomposition*, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 07), 2007.

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

[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 (2003), pp. 762–767.

[19] T. G. Kolda, B. W. Bader, and J. P. Kenny, *Higher-order web link analysis using multilinear algebra*, in Proc. of the 5th IEEE International Conference on Data Mining (ICDM 05), 2005.
[20] W. P. Krijnen, *Convergence of the sequence of parameters generated by alternating least squares algorithms*, Comput Statist. Data Anal., 51 (2006), pp. 481–489.
[21] L. D. Lathauwer, *Signal Processing based on Multilinear Algebra*, PhD thesis, Katholieke Universiteit Leuven, 1997.
[22] ———, *A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 642–666.
[23] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, *Blind source separation by simultaneous third-order tensor diagonalization*, in Proc. of the 8th European Signal Processing Conference (EUSIPCO 96), 1996.
[24] ———, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.
[25] ———, *On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1324–1342.
[26] ———, *Computation of the canonical decomposition by means of a simultaneous generalized schur decomposition*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 295–327.
[27] S. E. Leurgans, R. T. Ross, and R. B. Abel, *A decomposition for three-way arrays*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1064–1083.
[28] B. C. Mitchell and D. S. Burdick, *Slowly converging parafac sequences: Swamps and two-factor degeneracies*, J. Chemometrics, 8 (1994), pp. 155–168.
[29] P. Paatero, *A weighted non-negative least squares algorithm for three-way 'PARAFAC' factor analysis*, Chemometrics Intell. Lab. Syst., 38 (1997), pp. 223–242.
[30] A. Shashua and T. Hazan, *Non-negative tensor factorization with applications to statistics and computer vision*, in Proc. of the 22nd international conference on machine learning (ICML 05), 2005.
[31] V. D. Silva and L.-H. Lim, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM J. Matrix Anal. Appl., (to appear).
[32] A. Smilde, R. Bro, and P. Geladi, *Multi-way Analysis: Applications in the Chemical Sciences*, Wiley, 2004.
[33] J. Sun and C. Chen, *Generalized polar decomposition*, Math. Numer. Sin., 11 (1989), pp. 262–273.
[34] L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311.
[35] M. A. O. Vasilescu and D. Terzopoulos, *Multilinear subspace analysis for image ensembles*, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 03), 2003.
[36] H. Wang, Q. Wu, L. Shi, Y. Yu, and N. Ahuja, *Out-of-core tensor approximation of multi-dimensional matrices of visual data*, ACM Trans. Gr., 24 (2005), pp. 527–535.
[37] M. Welling and M. Weber, *Positive tensor factorization*, Pattern Recognition Letters, 22 (2001), pp. 1255–1261.
[38] T. Zhang and G. H. Golub, *Rank-one approximation to high order tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 534–550.