SIAM J. Matrix Anal. Appl.
Vol. 0, No. 0, pp. 000–000

ON THE GLOBAL CONVERGENCE OF THE ALTERNATING LEAST SQUARES METHOD FOR RANK-ONE APPROXIMATION TO GENERIC TENSORS*

LIQI WANG† AND MOODY T. CHU‡

Abstract. Tensor decomposition has important applications in various disciplines, but it remains an extremely challenging task even to this date. A slightly more manageable endeavor has been to find a low rank approximation in place of the decomposition. Even for this less stringent undertaking, it is an established fact that tensors beyond matrices can fail to have best low rank approximations, with the notable exception that the best rank-one approximation always exists for tensors of any order. Toward the latter, the most popular approach is the notion of alternating least squares whose specific numerical scheme appears in the form as a variant of the power method. Though the limiting behavior of the objective values is well understood, a proof of global convergence for the iterates themselves has been elusive. This paper partially addresses the missing piece by showing that for almost all tensors, the iterates generated by the alternating least squares method for the rank-one approximation converge globally. The underlying technique employed is an eclectic mix of knowledge from algebraic geometry and dynamical system.

Key words. tensor decomposition, rank-one tensor approximation, alternating least squares, global convergence, Zariski open set

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

DOI. 10.1137/130938207

1. Introduction.

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

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

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

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

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. By a tensor decomposition we refer to the rewriting of the given tensor T as the sum of some rank-one tensors, which can be regarded as a generalization of the singular value decomposition of matrices. Depending on the desirable structure, the two most prevalent summation formats are the general Tucker decomposition [9, 10, 33]

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

*Received by the editors September 23, 2013; accepted for publication (in revised form) by M. E. Kilmer June 4, 2014; published electronically DATE. This research was supported in part by the National Science Foundation under grant DMS-1014666.

http://www.siam.org/journals/simax/x-x/93820.html

†School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, 116024, People's Republic of China (lizzy@mail.dlut.edu.cn).

‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (chu@math.ncsu.edu).

and the "diagonal" case known as the CANDECOMP/PARAFAC decomposition [12, 15, 17, 19]

$$T = \sum_j \mathbf{u}_j^{(1)} \otimes \cdots \otimes \mathbf{u}_j^{(k)}. \tag{1.2}$$

Any of these decompositions finds a wide range of applications. Tensor decomposition has been a major ongoing research topic with numerous results available in the literature. No attempt will be made in this short note to detail these developments except by saying that decomposing a tensor amounts to finding real-valued solutions for a high-degree polynomial system of equations, which certainly is not a trivial task.

Instead of requiring a complete decomposition, an alternative approach attracting considerable interest in the field has been the low rank approximation. The basic setup consists of two parts. First, the number of terms in the summation is to be fixed a priori, whose proper choice is a difficult problem in itself [6, 21, 23]. Second, 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 is computationally more feasible. Nevertheless, it must be stressed that, except for the case of order-2 tensors which are simply matrices and the case of rank-one approximation for general tensors, the best low rank approximation for high-order tensors may not exist at all [11, 21, 24].

This paper concerns the rank-one approximation only. Our main contribution is proving the global convergence of the so called high-order power method (HOPM) for rank-one approximation. At first glance, this HOPM is simply an arithmetic analogy of the classical Gauss–Seidel method, but its underlying mathematic is equivalent to the alternating least squares (ALS) algorithm. Although the ALS method has been regarded as the most prevalent technique for low rank tensor approximation, the proof of global convergence of its iterates from any starting point even for the rank-one case has been elusive thus far [28, 34, 35]. Our goal is to partially close up that gap.

The chain of our arguments can be summarized as follows, which will be established step by step in the subsequent discussion:

1. The sequence produced by the algorithm has cluster points since it is bounded.
2. The sequence of increments tends to zero.
3. Cluster points are geometrically isolated for generic tensors.[1]
4. Together, items 2 and 3 imply that there is a unique limit point of the iteration per starting point.

So that the presentation is self-contained, we begin in section 2 with an introduction of our notation system and a brief review of some basic concepts. Item 2 is already established in [28], but we give a more specific estimate in section 3.1. The crux of our proof is the isolation of stationary points [28, section 4.3.16]. We will discuss how this concept can be motivated by the Łojasiewicz inequality for gradient flows in section 3.2 and employ an algebraic geometry argument in section 5 to establish item 3.

We should make it clear that while we will show in this paper that the ALS method does converge generically, still we have not answered the question of whether the ALS method might fail to converge at times. To settle this question we need

[1] Recall that a property holds almost everywhere if the set of elements for which the property does not hold is a set of Lebesque measure zero. Likewise, a generic tensor in our context means that those which do not satisfy the properties being described form a closed nowhere dense subset.

either an example showing the failure or a proof that a failure never occurs, which remains an open question at present.

2. Preliminaries.

This section contains some basic information of this subject and introduces the notation system used in this paper. It is particularly useful to recast the notion of "adjoint" by regarding tensors as linear operators on tensors. We also set up the basic HOPM scheme for later reference. Expert readers may skip reading through these elementary properties.

2.1. Linear operator.

An order-k tensor $T \in \mathbb{R}^{I_1\times\cdots\times I_k}$ may be thought of as a linear operator mapping order-$(k-1)$ tensors to vectors. Because there are multiple "facets," it should be specified that with respect to which facet of the tensor T that the dimension contraction is taking place. For this matter, we identify T as the operator $\mathscr{T} \equiv \mathscr{T}_\ell$, $\ell = 1, \ldots, k$, with

$$\mathscr{T}_\ell : \mathbb{R}^{I_1\times\cdots\times\widehat{I}_\ell\times\cdots\times I_k} \to \mathbb{R}^{I_\ell}, \tag{2.1}$$

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

$$\mathscr{T}_\ell(S) := T \circledast_\ell S = [\langle \tau_{:,\nu_\ell,:}, S\rangle] \in \mathbb{R}^{I_\ell}, \quad \nu_\ell = 1, \ldots, I_\ell, \tag{2.2}$$

where $\tau_{:,\nu_\ell,:}$ denotes the ν_ℓth "slice" of the tensor T in the ℓth direction and $\langle\cdot,\cdot\rangle$ is the Frobenius inner product generalized to multidimensional arrays. We remark that the operation $\circledast_\ell$, a natural result from operator perspective, is a convenient notation equivalent to multiple levels of contraction products [7] or n-mode products [21]. It can be described simply as the inner product of the tensor S with all subtensors of T extracted along the ℓth mode.

Taking an order-3 tensor acting on matrices as an example, let $\{E_J\}$, where J is a double index, and denote the standard basis $E_J = [e_{st}^J] \in \mathbb{R}^{m\times n}$ with

$$e_{st}^J = \begin{cases} 1 & \text{if } J = (s,t), \\ 0 & \text{otherwise,} \end{cases}$$

and likewise $\{\mathbf{e}_i\}$ the standard basis of $\mathbb{R}^p$. Suppose that the action of $\mathscr{T}$ at a basis matrix $E_J \in \mathbb{R}^{m\times n}$ is expressed in terms of basis $\{\mathbf{e}_i\}$ as

$$\mathscr{T}(E_J) = \sum_i \tau_{i,J}\mathbf{e}_i$$

with $\tau_{i,J} \in \mathbb{R}$. Then for any $X = [x_{ij}] \in \mathbb{R}^{m\times n}$, we have the relationship

$$\mathscr{T}(X) = \mathscr{T}\left(\sum_J x_J E_J\right) = \sum_J x_J \left(\sum_i \tau_{i,J} e_i\right) = \sum_i \left(\sum_J \tau_{i,J} x_J\right)\mathbf{e}_i.$$

The order-3 tensor T may be visualized as a $p \times 1$ block matrix of which each block is an $m \times n$ matrix, whereas the action $\circledast_1$ is a two-dimensional contraction defined by

$$\mathscr{T}(X) \equiv \underbrace{\begin{bmatrix} \boxed{\tau_{1,:}} \\ \boxed{\tau_{2,:}} \\ \vdots \\ \boxed{\tau_{p,:}} \end{bmatrix}}_{T} \circledast_1 X := [\langle \tau_{i,:}, X\rangle] \in \mathbb{R}^p.$$

It follows naturally that the adjoint operator $\mathscr{T}^*:\mathbb{R}^p\rightarrow\mathbb{R}^{m\times n}$ should be defined as

$$\mathscr{T}^*(\mathbf{v}) := \sum_i v_i\tau_i, \tag{2.3}$$

which satisfies the requirement that

$$\langle\mathscr{T}(X),\mathbf{v}\rangle = \langle X,\mathscr{T}^*(\mathbf{v})\rangle \tag{2.4}$$

for every $X\in\mathbb{R}^{m\times n}$ and $\mathbf{v}\in\mathbb{R}^p$. Generalizing the spirit of (2.4), whereas the notion of "transpose" is embedded in the tensor product $\circledast_\ell$, the following result can easily be proved by rearranging terms in the summation via 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*

$$\left\langle T,\mathbf{u}^{(1)}\otimes\cdots\otimes\mathbf{h}^{(\ell)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right\rangle = \left\langle T\circledast_\ell\left(\mathbf{u}^{(1)}\otimes\cdots\otimes\widehat{\mathbf{u}}^{(\ell)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right),\mathbf{h}^{(\ell)}\right\rangle. \tag{2.5}$$

Because (2.5) holds for any index ℓ, we have the reciprocating relationship

$$\left\langle T\circledast_i\left(\mathbf{u}^{(1)}\otimes\cdots\otimes\widehat{\mathbf{u}}^{(i)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right),\mathbf{u}^{(i)}\right\rangle = \left\langle T\circledast_j\left(\mathbf{u}^{(1)}\otimes\cdots\otimes\widehat{\mathbf{u}}^{(j)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right),\mathbf{u}^{(j)}\right\rangle \tag{2.6}$$

for any two distinct indices i and j. Both identities will be extremely useful in the subsequent discussion.

Obviously, the linearity of T can be interpreted more generally. For instance, if $k\geq 3$ and $i<j$, then we can also interpret T as the representation of the linear operator

$$\mathscr{T}_{i,j}:\mathbb{R}^{I_1\times\cdots\times\widehat{I}_i\times\cdots\times\widehat{I}_j\times\cdots\times I_k}\rightarrow\mathbb{R}^{I_i\times I_j}, \tag{2.7}$$

via the action

$$\mathscr{T}_{i,j}(U) := T\circledast_{i,j}U = \left[\langle\tau_{:,\nu_i,:,\nu_j,:},U\rangle\right]\in\mathbb{R}^{I_i\times I_j},\quad \nu_i=1,\ldots,I_i;\ \nu_j=1,\ldots,I_j, \tag{2.8}$$

for any order-$(k-2)$ tensor $U\in\mathbb{R}^{I_1\times\cdots\times\widehat{I}_i\times\cdots\times\widehat{I}_j\times\cdots\times I_k}$.

2.2. HOPM. The problem of finding a best rank-one approximation to T is to determine unit vectors $\mathbf{u}^{(j)}\in\mathbb{R}^{I_j}$, $j=1,\ldots k$ and a scalar λ such that the function

$$\begin{aligned} f\left(\lambda,\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(k)}\right) &:= \left\|T-\lambda\mathbf{u}^{(1)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right\|_F^2 \\ &= \sum_{i_1,i_2,\ldots,i_k}\left(\tau_{i_1,\ldots,i_k}-\lambda u_{i_1}^{(1)}\ldots u_{i_k}^{(k)}\right)^2 \end{aligned} \tag{2.9}$$

is minimized. It is not difficult to see that for any fixed unit vectors $\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(k)}$, the optimal value of λ for (2.9) is necessarily given by

$$\lambda = \lambda\left(\mathbf{u}^{(1)},\ldots,\mathbf{u}^{(k)}\right) = \left\langle T,\mathbf{u}^{(1)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right\rangle. \tag{2.10}$$

To put it differently, by regarding the order-k tensor space as a vector space equipped with the Frobenius inner product, the expression for λ in (2.10) is precisely the length of the projection of the "vector" T onto the direction of the "unit vector" $\mathbf{u}^{(1)}\otimes\cdots\otimes\mathbf{u}^{(k)}$. Thus, minimizing the orthogonal component of T, as is so desired in minimizing the cost function (2.9), is equivalent to maximizing the parallel component λ. In [35], the expression (2.10) is called the generalized Rayleigh quotient.

Using the Lagrange multiplier theory, we can easily argue, with the aid of (2.5), that the first order optimality condition for a stationary point is to satisfy the system of equations [25]

$$T \circledast_\ell \left(\mathbf{u}^{(1)} \otimes \cdots \otimes \widehat{\mathbf{u}}^{(\ell)} \otimes \cdots \otimes \mathbf{u}^{(k)}\right) = \lambda \mathbf{u}^{(\ell)}, \quad \ell = 1, \ldots, k. \tag{2.11}$$

To solve (2.11), a commonly used tactic is to alternate directions among the variables in a way similar to the classical Gauss–Seidel method [13]. See Algorithm 1 for the description, where the subscript $\cdot_{[p]}$ indicates the quantities resulting from the pth iteration.

Algorithm 1. (HOPM.)

```
Given $\mathbf{u}^{(1)}_{[0]} \in \mathbb{R}^{I_1}, \ldots, \mathbf{u}^{(k)}_{[0]} \in \mathbb{R}^{I_k}$,
for $p = 0, 1, \ldots,$ do
  for $\ell = 1, 2, \ldots, k$ do
    $\mathbf{u}^{(\ell)}_{[p+1]} = T \circledast_\ell \left(\mathbf{u}^{(1)}_{[p+1]} \otimes \cdots \otimes \mathbf{u}^{(\ell-1)}_{[p+1]} \otimes \widehat{\mathbf{u}}^{(\ell)} \otimes \mathbf{u}^{(\ell+1)}_{[p]} \cdots \otimes \mathbf{u}^{(k)}_{[p]}\right)$
    $\lambda^{(\ell)}_{[p+1]} := \|\mathbf{u}^{(\ell)}_{[p+1]}\|_2$
    $\mathbf{u}^{(\ell)}_{[p+1]} := \frac{\mathbf{u}^{(\ell)}_{[p+1]}}{\lambda^{(\ell)}_{[p+1]}}$
  end for
end for
```

In the case of $k = 2$, the HOPM generates normal vectors $\{\mathbf{u}^{(1)}_{[p]}\}$ and $\{\mathbf{u}^{(2)}_{[p]}\}$ which satisfy

$$\begin{cases} \mathbf{u}^{(1)}_{[p+1]} = \frac{1}{\lambda^{(1)}_{[p+1]}\lambda^{(2)}_{[p]}} TT^\top \mathbf{u}^{(1)}_{[p]}, \\ \mathbf{u}^{(2)}_{[p+1]} = \frac{1}{\lambda^{(1)}_{[p+1]}\lambda^{(2)}_{[p+1]}} T^\top T \mathbf{u}^{(2)}_{[p]} \end{cases} \tag{2.12}$$

and are precisely the iterates obtained by applying the classical power method to the matrices $TT^\top$ and $T^\top T$, respectively. As such, the convergence of sequences $\{\mathbf{u}^{(1)}_{[p]}\}$ and $\{\mathbf{u}^{(2)}_{[p]}\}$ in (2.12) is well understood, and the limit point of either $\{\lambda^{(1)}_{[p]}\}$ or $\{\lambda^{(2)}_{[p]}\}$ is precisely the largest singular value of T. When the HOPM is applied to high-order tensors, however, the global convergence is an interesting but only partially resolved area of research [28, 34, 35]. A proof of global convergence of the iterates, even for rank-one approximation, would be significant, and that is precisely the primary purpose of this paper.

2.3. ALS method. The Gauss–Seidel type iteration in the HOPM has an inherent optimization property—it can be regarded as an ALS method [2, 15] applied to (2.10). It has an ascending property in the following sense. For a given p and at the ℓth stage, consider the maximization of the function $\theta_\ell : \mathbb{R}^{I_\ell} \to \mathbb{R}$ defined by

$$\theta_\ell\left(\mathbf{y}^{(\ell)}\right) := \left\langle T, \mathbf{u}^{(1)}_{[p+1]} \otimes \cdots \otimes \mathbf{u}^{(\ell-1)}_{[p+1]} \otimes \mathbf{y}^{(\ell)} \otimes \mathbf{u}^{(\ell+1)}_{[p]} \otimes \cdots \otimes \mathbf{u}^{(k)}_{[p]}\right\rangle, \tag{2.13}$$

subject to the constraint $\|\mathbf{y}^{(\ell)}\|_2 = 1$. It immediately follows from (2.5) and the Cauchy–Schwartz inequality that θ_ℓ is maximized at

$$\mathbf{y}^{(\ell)} = \mathbf{u}^{(\ell)}_{[p+1]} := \frac{T \circledast_\ell \left(\mathbf{u}^{(1)}_{[p+1]} \otimes \cdots \otimes \mathbf{u}^{(\ell-1)}_{[p+1]} \otimes \widehat{\mathbf{u}}^{(\ell)} \otimes \mathbf{u}^{(\ell+1)}_{[p]} \cdots \otimes \mathbf{u}^{(k)}_{[p]}\right)}{\lambda^{(\ell)}_{[p+1]}},$$

provided that the denominator $\lambda^{(\ell)}_{[p+1]}$ is not zero. By repeating this argument to successive ℓ, we conclude that

(2.14)
$$\lambda_{[p]} := \lambda\left(\mathbf{u}^{(1)}_{[p]}, \ldots, \mathbf{u}^{(k)}_{[p]}\right) \leq \lambda^{(1)}_{[p+1]} \leq \lambda^{(2)}_{[p+1]} \leq \cdots \leq \lambda^{(k)}_{[p+1]} = \lambda\left(\mathbf{u}^{(1)}_{[p+1]}, \ldots, \mathbf{u}^{(k)}_{[p+1]}\right).$$

In other words, so long as the starting points $\mathbf{u}^{(\ell)}_{[0]}$, $\ell = 1, \ldots, k$, are such that $\lambda^{(1)}_{[1]} \neq 0$, each sweep in the inner loop of Algorithm 1 has the effect of increasing the functional value of λ. Because the range of λ is bounded by $\|T\|_F$, the sequence $\left\{\lambda_{[p]}\right\}$ must converge.

The ALS method formulated in Algorithm 1 is for rank-one approximation only. Its present form as a (high-order) power method requires no inversion and makes the computation straightforward. The idea can be generalized to higher rank approximation in various ways and has been used as the "workhorse" algorithm for computing rank decomposition of general tensors [6, 21].

The preceding argument about the nondecreasing property of $\lambda_{[p]}$ is essentially the only evidence of convergence in the literature [20, 35]. This certainly is not enough, because convergence of objective values does not guarantee the convergence of the iterates themselves [21]. This lack of proof in general is recognized by researchers as, for example, it was claimed in [8] that "Their [the ALS algorithms] extensive use is thus unexplainable." See also the many musings in the paper [28, section 4]. More needs be done. We offer the following convergence analysis as only one small step toward that goal.

3. Global convergence.

The first formal proof of linear convergence of the ALS iterates was probably made in [35, Theorem 4.3]. The argument there, for rank-one approximation only, was based on a comparison with the Newton method. The same question of linear convergence for higher-rank approximations by the ALS was considered recently in [34] with a more accessible argument. In both cases, however, the proof is for local convergence of the iterates. We hereby present a proof of global convergence of the ALS iterates for rank-one approximation. The target tensor can be of arbitrary order.

Additionally, we stress that our global convergence proof is unconditional and generic, which is not the case for [34], where the assumption is that the Hessian matrix of the underlying problem is positive definite modulo the scaling indeterminacy, nor for [35], where some special rank conditions on the objective must be imposed. We shall not concern ourselves with the rate of convergence, though it is expected to be linear. Rather, our focus is on the global convergence of the iterates.

3.1. Diminishing increments.

Concerning the convergence of the ALS applied to general multilinear functions, an array of critical questions has been raised and partially answered in the interesting paper [28]. For our application, the following result can be found in [28, section 4.3.2].

Lemma 3.1. *The sequence* $\{\mathbf{u}_{[p]}^{(1)}\otimes\cdots\otimes\mathbf{u}_{[p]}^{(k)}\}$ *of rank-one tensors generated by the HOPM satisfies*

$$\left\|\mathbf{u}_{[p+1]}^{(1)}\otimes\cdots\otimes\mathbf{u}_{[p+1]}^{(k)}-\mathbf{u}_{[p]}^{(1)}\otimes\cdots\otimes\mathbf{u}_{[p]}^{(k)}\right\|\to 0 \tag{3.1}$$

as p goes to infinity.

Note that rank-one tensor approximant consists of one single term. Since it involves no summation, there is no possibility of cancelation by other terms. Thus, the same proof in [28] for Lemma 3.1 can be employed to show that for any fixed p, the increments among factors, that is,

$$\delta\mathbf{u}_{[p]}^{(\ell)}:=\mathbf{u}_{[p+1]}^{(\ell)}-\mathbf{u}_{[p]}^{(\ell)},\quad \ell=1,\ldots,k, \tag{3.2}$$

also converge to zero as p goes to infinity.

Indeed, it might be more informative to specifically quantify the magnitude of $\delta\mathbf{u}_{[p]}^{(\ell)}$ as follows. First, the identities

$$\left\langle\mathbf{u}_{[p+1]}^{(\ell)},\delta\mathbf{u}_{[p]}^{(\ell)}\right\rangle=1-\left\langle\mathbf{u}_{[p+1]}^{(\ell)},\mathbf{u}_{[p]}^{(\ell)}\right\rangle=\frac{1}{2}\left\|\delta\mathbf{u}_{[p]}^{(\ell)}\right\|^2 \tag{3.3}$$

are trivial to verify. Then, using (3.2) and denoting $\lambda_{[p+1]}^{(0)}:=\lambda_{[p]}$, we can also establish the identity that

$$\begin{aligned}\lambda_{[p+1]}^{(\ell)}&=\theta_\ell\left(\mathbf{u}_{[p+1]}^{(\ell)}\right)=\left\langle T,\mathbf{u}_{[p+1]}^{(1)}\otimes\cdots\otimes\mathbf{u}_{[p+1]}^{(\ell-1)}\otimes\left(\mathbf{u}_{[p]}^{(\ell)}+\delta\mathbf{u}_{[p]}^{(\ell)}\right)\otimes\mathbf{u}_{[p]}^{(\ell+1)}\otimes\cdots\otimes\mathbf{u}_{[p]}^{(k)}\right\rangle\\&=\lambda_{[p+1]}^{(\ell-1)}+\left\langle\delta\mathbf{u}_{[p]}^{(\ell)},\underbrace{T\circledast_\ell\,\mathbf{u}_{[p+1]}^{(1)}\otimes\cdots\otimes\mathbf{u}_{[p+1]}^{(\ell-1)}\otimes\widehat{\mathbf{u}}^{(\ell)}\otimes\mathbf{u}_{[p]}^{(\ell+1)}\cdots\otimes\mathbf{u}_{[p]}^{(k)}}_{\lambda_{[p+1]}^{(\ell)}\mathbf{u}_{[p+1]}^{(\ell)}}\right\rangle\\&=\lambda_{[p+1]}^{(\ell-1)}+\frac{\lambda_{[p+1]}^{(\ell)}}{2}\left\|\delta\mathbf{u}_{[p]}^{(\ell)}\right\|^2,\end{aligned} \tag{3.4}$$

which sheds additional insight to the inequalities (2.14). Note that since all terms in (3.4) are nonnegative, we must have $\|\delta\mathbf{u}_{[p]}^{(\ell)}\|\leq\sqrt{2}$, where the equality holds only if $\lambda_{[p+1]}^{(\ell-1)}=0$. The relationship (3.4) is analogous to the known fact of quadratic convergence of the Rayleigh quotients with respect to eigenvector improvement.

The convergence of $\{\delta\mathbf{u}_{[p]}^{(\ell)}\}$ only shows that the gap between two consecutive iterates $\mathbf{u}_{[p]}^{(\ell)}$ and $\mathbf{u}_{[p+1]}^{(\ell)}$ is getting smaller, but does not guarantee that $\{\mathbf{u}_{[p]}^{(\ell)}\}$ is forming a Cauchy sequence as is required of its convergence. This missing component from the global convergence analysis has long been recognized in the literature [28, section 4.3.3] and is precisely where our contribution in this paper becomes useful.

3.2. Geometrically isolated stationary points. The primary tool we propose to use is motivated from the notion of gradient flows. So that this paper is self-contained, we briefly review some basic principles for readers who might not be familiar with this dynamical system.

We begin with the classical Łojasiewicz inequality [3, 26, 27], which has a wide range of applications.

Theorem 3.2. *Suppose that $F:U\to\mathbb{R}$ is real analytic in an open set $U\subset\mathbb{R}^n$. Then for any point $\mathbf{p}\in U$, there exist a neighborhood W of $\mathbf{p}$, constants $\theta\in[\frac{1}{2},1)$,* *and* $c > 0$ *such that*

$$\|F(\mathbf{x}) - F(\mathbf{p})\|^{\theta} \leq c\|\nabla F(\mathbf{x})\| \quad \textit{for all } \mathbf{x} \in W. \tag{3.5}$$

Consider the gradient flow

$$\frac{d\mathbf{x}}{dt} = -\nabla F(\mathbf{x}) \tag{3.6}$$

for minimizing an objective function $F(\mathbf{x})$, where the derivative $\frac{d\mathbf{x}}{dt}$ of the path $\mathbf{x}(t)$ is taken with respect to some parameter t. It is known that if $\mathbf{x}(t)$ is a bounded semiorbit of (3.6), even if F is merely differentiable, then the set of accumulation points, that is,

$$\omega(\mathbf{x}(0)) := \{\mathbf{x}^* \in \mathbb{R}^n \mid \mathbf{x}(t_\nu) \to \mathbf{x}^* \text{ for some sequence } t_\nu \to \infty\} \tag{3.7}$$

is a nonempty, compact, and connected subset of stationary points

$$\mathcal{C} := \{\mathbf{x} \in \mathbb{R}^n \mid \nabla F(\mathbf{x}) = 0\}. \tag{3.8}$$

For analytic gradient flow, the limiting behavior in (3.7) is considerably stronger. The next theorem asserts that the set $\omega(\mathbf{x}(0))$ of any analytic gradient flow $\mathbf{x}(t)$ is necessarily a singleton. This fact, a consequence of the Łojasiewicz inequality, is of fundamental importance in gradient flow applications. We recommend the reference [1, Theorem 2.2] and the lecture note [29] for its proof.

THEOREM 3.3. *Suppose that* $F : U \to \mathbb{R}$ *is real analytic in an open set* $U \subset \mathbb{R}^n$. *Then for any bounded semiorbit of* (3.6), *there exists a point* $\mathbf{x}^* \in \mathcal{S}$ *such that* $\mathbf{x}(t) \to \mathbf{x}^*$ *as* $t \to \infty$.

Consider the sequence $\{(\mathbf{u}_{[p]}^{(1)}, \dots, \mathbf{u}_{[p]}^{(k)})\}$ generated by the ALS iteration. It is obvious that any accumulation point of this sequence necessarily satisfies the system of nonlinear equations (2.11), whereas the system itself characterizes an analytic gradient field. The latter can be seen from the fact the (partial) gradient of λ with respect to $\mathbf{u}^{(\ell)}$ is given by

$$\nabla^{(\ell)}\lambda := T \circledast_\ell \left(\mathbf{u}^{(1)}\otimes\mathbf{u}^{(2)}\otimes\cdots\otimes\widehat{\mathbf{h}}^{(\ell)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right), \quad \ell = 1, \dots k. \tag{3.9}$$

Because of the constraints $\|\mathbf{u}^{(\ell)}\| = 1$, the projected gradient is given by

$$\mathbf{Proj}_{\|\mathbf{u}^{(\ell)}\|=1}\nabla^{(\ell)}\phi = \nabla^{(\ell)}\phi - \left\langle \nabla^{(\ell)}\phi, \mathbf{u}^{(\ell)} \right\rangle \mathbf{u}^{(\ell)}, \quad \ell = 1, \dots, k. \tag{3.10}$$

So the dynamical system

$$\frac{d\mathbf{u}^{(\ell)}}{dt} = \underbrace{T \circledast_\ell \left(\mathbf{u}^{(1)}\otimes\mathbf{u}^{(2)}\otimes\cdots\otimes\widehat{\mathbf{h}}^{(\ell)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right) - \lambda\mathbf{u}^{(\ell)}}_{\nabla^{(\ell)}F\left(\mathbf{u}^{(1)},\dots,\mathbf{u}^{(k)}\right)}, \quad \ell = 1, \dots, \dots, k, \tag{3.11}$$

defines an ascending gradient flow $\mathbf{u}^{(\ell)}(t)$ on the unit ball for the function λ. By Theorem 3.3, the solution flow $\mathbf{u}^{(\ell)}(t)$ must converge to a single limit point. This dynamical behavior for the differential system (3.11) should have addressed the concern raised in [28, section 4.3.16], namely, it was not clear how to exclude the possibility that the tensor approximation produced by the ALS algorithm could fail to converge in a bounded way, such as spiraling in toward a limit circle or cluster points. For our application, since the ALS method is a discrete dynamical system, we need something stronger on the topology of the stationary points. The following theorem will be proved in the appendix, where we shall explain in further detail the notion that the set of tensor T satisfying the theorem forms a Zariski open set[2] [30].

THEOREM 3.4. *For almost all tensors $T \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_k}$, the set $\mathcal{C}$ of points $(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)})$, where $\nabla^{(\ell)} F(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}) = 0$, $\ell = 1, \ldots, k$, contains only geometrically isolated points.*

3.3. Componentwise global convergence. To prove the global convergence of the ALS iterates, we use a tool modified from a result proved in [5, Lemma 4.3].

LEMMA 3.5. *Suppose that $\{a_n\}$ is a bounded sequence of real numbers where the difference $|a_{n+1} - a_n|$ converges to zero as n goes to infinity. If the accumulation points for the sequence are geometrically isolated, then the sequence $\{a_n\}$ must converge 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

$$\begin{cases} \quad a_{\alpha_k} \in B_x(\epsilon), \\ \quad a_{\beta_k} \in B_y(\epsilon), \\ |a_{k+1} - a_k| < \epsilon \end{cases}$$

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 an accumulation point. Since z is arbitrary, any number in between x and y is an accumulation point. This contradicts the assumption that these accumulations points are geometrically isolated. □

Finally, we complete the last stage of convergence analysis as follows.

THEOREM 3.6. *For almost all tensors $T \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_k}$, the sequence $\{\mathbf{u}_{[p]}^{(\ell)}\}$ generated by the ALS method converges for each $\ell = 1, \ldots, k$.*

Proof. Starting with any initial value $\mathbf{u}_{[0]}^{(\ell)}$, it is clear that the iterates $\{\mathbf{u}_{[p]}^{(\ell)}\}$ are bounded. By definition, any accumulation points necessarily satisfy the system of nonlinear equations (2.11) whose solutions, by Theorem 3.3, are geometrically isolated. We already know $\{\delta \mathbf{u}_{[p]}^{(\ell)}\}$ converges to zero as p goes to infinity. It follows that the sequence $\{\mathbf{u}_{[p]}^{(\ell)}\}$ converges to a single limit point because, by Lemma 3.5, each component converges. □

The argument advanced above for the global convergence of the ALS method on the rank-one tensor approximation represents only a small step toward the general theory of low rank tensor approximation, but it is perhaps also the first such proof in the literature thus far. A proof of orthogonal rank-2 tensor approximation will be given in a separate paper.

3.4. Local optimizers. Though we have established the global convergence of the ALS method, it must be stressed that the objective function f defined in (2.9) generally has multiple local minimizers (and thus multiple local maximizers for λ). A limit point obtained by the ALS method depends on the starting point and is not

[2] In the simplest term, the set of common zeroes of a collection of polynomials is called a Zariski closed set. The complement of a Zariski closed set is a Zariski open set.

Table 1

Percentage of numbers of limit points found by the ALS method for 200 *random tensors in* $\mathbb{R}^{10\times 5\times 3}$.

# of limit points	1	2	3	4
Percentage of occurrence	29.50%	52.00%	17.00%	1.50%

guaranteed to be the best rank-one approximation to T [32]. Questions such as about the number of stationary points and starting strategies naturally arise.

For order-2 tensors, the answer to the first question is explicitly known. Stationary points to (2.9) are precisely those triplets of singular values and associated singular vectors of the matrix T. Furthermore, the minimizer of the functional f (or maximizer of λ) is typically unique, which is the triplet corresponding to the largest singular value, whereas all other triplets are saddle points. But for tensors of higher order, the answer is indefinite. Taking 200 random tensors with i.i.d. Gaussian distribution in $\mathbb{R}^{10\times 5\times 3}$, and applying the ALS method with 20 random initial values to each given tensor, for example, we find that there are one to four limit points with a distribution of numbers of limit points depicted in Table 1. Since the ALS algorithm is an ascent method for the generalized Rayleigh quotient λ, these limit points are local minimizers for f. How the cardinalities of limit points vary in accordance with the size of the tensors is an open question. It clearly illustrates however that, when multiple limit points are present, there is a high possibility that the best rank-one approximation is not found. How to propose a suitable starting strategy so as to help to improve the performance of the ALS algorithm is also a topic that might deserve further investigation.

4. Conclusion.

The ALS algorithm has been regarded as the workhorse method for computing tensor rank decomposition and low rank approximations. Its application to rank-one approximation appears as a HOPM whose computation is quite straightforward. This paper offers a proof, probably for the first time, that the iterates generated by the ALS method converge globally for generic tensors. The approach we have adopted is a synthesis of tools from linear theory, dynamical system, and algebraic geometry.

It is important to point out that, while we have shown that the ALS method does converge generically, we have not answered the question of whether the ALS method might fail to converge at times. To settle this question, we need either an example showing the failure or a proof that a failure never occurs.

For high-order tensors, the best low rank approximation may not exist. Also, the ALS scheme will necessarily involve inversions when solving the resulting linear systems. Under specific conditions, such as mutually orthogonal rank-one components [22, 31], the theory developed in this paper can be generalized and global convergence can also be ensured. Results of this investigation will be reported in a separate paper.

5. Appendix: Proof of Theorem 3.4.

To prove that a stationary point, that is, a root of the projected gradient, is geometrically isolated, it suffices to prove that the Jacobian of the system $\nabla^{(\ell)}F$, $\ell = 1,\ldots,k$, is of full rank at such a point [18, section 2]. The key in proving Theorem 3.4 is to characterize the peculiar tensor T at which the corresponding Jacobian is singular.

5.1. Order-2 tensors.

It will be instructive to first consider the case of $k = 2$, which should shed light on the general case. Given a matrix $T \in \mathbb{R}^{m\times n}$, the corresponding system of equations (2.11) can be written as

$$\left\{\begin{array}{l} T\mathbf{v} = \sigma\mathbf{u}, \\ T^\top\mathbf{u} = \sigma\mathbf{v} \end{array}\right. \tag{5.1}$$

with $\sigma = \mathbf{u}^\top T\mathbf{v}$. The nontrivial solutions therefore are the singular triplets $(\sigma_i, \mathbf{u}_i, \mathbf{v}_i)$, $i = 1, \ldots r$, from the singular value decomposition

$$T = U\Sigma V^\top$$

of the matrix T, where r denotes the rank of T. The Jacobian matrix $J(\mathbf{u}, \mathbf{v})$ at such a point, say, $(\sigma_j, \mathbf{u}_j, \mathbf{v}_j)$, is given by

$$J(\mathbf{u}_j, \mathbf{v}_j) = \left[\begin{array}{cc} -\sigma_j(I_m + \mathbf{u}_j\mathbf{u}_j^\top) & T(I_n - \mathbf{v}_j\mathbf{v}_j^\top) \\ T^\top(I_m - \mathbf{u}_j\mathbf{u}_j^\top) & -\sigma_j(I_n + \mathbf{v}_j\mathbf{v}_j^\top) \end{array}\right]. \tag{5.2}$$

Since $\sigma_j > 0$, the $(1, 1)$ block of $J(\mathbf{u}_j, \mathbf{v}_j)$ is negative definite. Upon eliminating the $(2, 1)$ block, the Schur complement at the $(2, 2)$ block is given by

$$V\mathrm{diag}\left\{\frac{\sigma_1^2}{\sigma_j} - \sigma_j, \ldots, -2\sigma_j, \ldots \frac{\sigma_r^2}{\sigma_j} - \sigma_j\right\} V^\top, \tag{5.3}$$

where the entry $-2\sigma_j$ occurs at the jth diagonal position. It follows that if σ_j is distinct from all other nonzero singular values of T, then $J(\mathbf{u}_j, \mathbf{v}_j)$ is nonsingular and, hence, $(\mathbf{u}_j, \mathbf{v}_j)$ is geometrically isolated. The condition that the matrix T has distinct nonzero singular values is a generic phenomenon, so Theorem 3.4 is true for the case of $k = 2$.

With (5.2) in hand, we can further characterize the global maximizer of λ. Since the system of equations (5.1) represents the projected gradient, Jacobian (5.2) applied to tangent vectors of the feasible set represents the project Hessian [4]. Suppose that the singular values of T are arranged in descending order $\sigma_1 \geq \sigma_2 \geq \sigma_p > \sigma_{p+1} = \cdots = 0$. Let $\widetilde{\mathbf{u}}_j$ and $\widetilde{\mathbf{v}}_j$ denote any of the other singular vectors, including those corresponding to zero singular values. These vectors are orthogonal to $\mathbf{u}_j$ and $\mathbf{v}_j$, respectively, and thus span the corresponding tangent spaces. Then

$$\left[\widetilde{\mathbf{u}}_j^\top, \widetilde{\mathbf{v}}_j^\top\right]^\top J(\mathbf{u}_j, \mathbf{v}_j) \left[\begin{array}{c} \widetilde{\mathbf{u}}_j \\ \widetilde{\mathbf{v}}_j \end{array}\right] = 2(\widetilde{\sigma} - \sigma_j),$$

where $\widetilde{\sigma} := \widetilde{\mathbf{u}}_j^\top T\widetilde{\mathbf{v}}_j$. In order that the projected Hessian is negative definite over the tangent space of $(\mathbf{u}_j, \mathbf{v}_j)$, it is now clear that it is possible only when $j = 1$, that is, $(\sigma_1, \mathbf{u}_1, \mathbf{v}_1)$ is the only local and global maximizer for λ. This view from the second order optimality condition offers an alternative explanation on the convergence of the iterates by the power method (2.12) to the dominant singular vectors.

5.2. The general case. The Jacobian matrix $J = J(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)})$ of the full system is a symmetric $k \times k$ block matrix whose (i, j) block is of size $I_i \times I_j$ and is given by

(5.4)

$$\frac{\partial\nabla^{(i)}F}{\partial\mathbf{u}^{(j)}} = \begin{cases} -\lambda\left(I_{I_i} + \mathbf{u}^{(i)}\mathbf{u}^{(i)^\top}\right) & \text{if } j = i, \\ T\circledast_{i,j}\left(\mathbf{u}^{(1)}\otimes\cdots\otimes\widehat{\mathbf{u}}^{(i)}\otimes\cdots\otimes\widehat{\mathbf{u}}^{(j)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right) - \lambda\mathbf{u}^{(i)}\mathbf{u}^{(j)^\top} & \text{if } i < j, \\ \left(T\circledast_{j,i}\left(\mathbf{u}^{(1)}\otimes\cdots\otimes\widehat{\mathbf{u}}^{(j)}\otimes\cdots\otimes\widehat{\mathbf{u}}^{(i)}\otimes\cdots\otimes\mathbf{u}^{(k)}\right)\right)^\top - \lambda\mathbf{u}^{(i)}\mathbf{u}^{(j)^\top} & \text{if } i > j, \end{cases}$$

where the operation $\circledast_{i,j}$ is defined earlier in (2.8). As T has multiple facets, the analysis of nonsingularity of the Jacobian matrix J is more complicated. However, the basic idea is the same as that for the case of $k = 2$. We explain the idea below.

Without loss of generality, let $W_{12} \in \mathbb{R}^{I_1 \times I_2}$ denote the matrix

$$W_{12} := T \circledast_{1,2} \left(\mathbf{u}^{(3)} \otimes \cdots \otimes \mathbf{u}^{(k)} \right). \tag{5.5}$$

Then, by definition, we have

$$\left\{ \begin{array}{l} W_{12} \circledast_1 \mathbf{u}^{(2)} = T \circledast_1 \left(\widehat{\mathbf{u}}^{(1)} \otimes \mathbf{u}^{(2)} \otimes \cdots \otimes \mathbf{u}^{(k)} \right) = \lambda \mathbf{u}^{(1)}, \\ W_{12} \circledast_2 \mathbf{u}^{(1)} = T \circledast_2 \left(\mathbf{u}^{(1)} \otimes \widehat{\mathbf{u}}^{(2)} \otimes \cdots \otimes \mathbf{u}^{(k)} \right) = \lambda \mathbf{u}^{(2)}. \end{array} \right. \tag{5.6}$$

In other words, whenever $(\lambda, \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)})$ is a root of the nonlinear system (2.11), the corresponding triplet $(\lambda, \mathbf{u}^{(1)}, \mathbf{u}^{(2)})$ for $\lambda > 0$ is a singular triplet for the matrix W_{12}. The upper-left 2×2 block of the Jacobian matrix J is simply of the same form as that of (5.2), that is,

$$J_{12}(\mathbf{u}_j^{(1)}, \mathbf{u}_j^{(2)}) = \left[\begin{array}{cc} -\lambda_j \left(I_m + \mathbf{u}_j^{(1)} \mathbf{u}_j^{(1)^\top} \right) & W_{12} \left(I_n - \mathbf{u}_j^{(2)} \mathbf{u}_j^{(2)^\top} \right) \\ W_{12}^\top \left(I_m - \mathbf{u}_j^{(1)} \mathbf{u}_j^{(1)^\top} \right) & -\lambda_j \left(I_n + \mathbf{u}_j^{(2)} \mathbf{u}_j^{(2)^\top} \right) \end{array} \right],$$

where the notation $(\lambda_j, \mathbf{u}_j^{(1)}, \mathbf{u}_j^{(2)})$ is meant to indicate one particular singular triplet. It follows from the analysis in section 5.1 that in order that the Schur compliment of the $(1, 1)$ block of J_{12} is nonsingular, it is necessary that W_{12} has distinct nonzero singular values. From the expression (5.6) and the explanation made earlier for (5.2), W_{12} has distinct singular values for all most all T. Continuing this block Gaussian elimination procedure and using the Woodbury identity

$$(A - UCV)^{-1} = A^{-1} + A^{-1}U(C^{-1} - VA^{-1}U)^{-1}VA^{-1}$$

to express the inverse of the Schur complement recursively, we should be able to formally transform the Jacobian matrix $J(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)})$ into a block diagonal matrix. The procedure breaks down only if one of the first $k - 1$ diagonal blocks (Schur compliments) is not invertible, which can happen only when T satisfies some stringent algebraic conditions. Finally, for the (k, k) diagonal block to be singular, T must be subject to additional constraints. In all, for the Jacobian matrix $J(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)})$ to be rank deficient, the entries of T must satisfy a collection of algebraic constraints whose zero locus forms an algebraic variety in the ambient space $\mathbb{R}^{I_1 \times \cdots \times I_k}$ [16]. The Sard theorem and the preimage theorem assert that this variety generically is lower dimensional, so it is of measure zero[3] [14]. That is, for almost all T, the Jacobian at the zeros of the projected gradient $\nabla^{(\ell)} F$, $\ell = 1, \ldots k$ is nonsingular and, thus, these stationary points are geometrically isolated.

5.3. Algebraic geometry viewpoint. What we have argued above is to examine the determinant of the Jacobian matrix directly by block Gaussian elimination. The same result can be obtained more concisely in the context of algebraic geometry. We briefly outline the idea in this section.

The system of nonlinear equations (2.11) can be regarded as a polynomial system whose "leading" terms, upon replacing λ by (2.10), are homogeneous of degree

[3] A stronger statement, which is applicable in our case, is that if a real analytic function $f : U \subset \mathbb{R}^n \to \mathbb{R}^m$ is zero on a set Z of positive measure (and U is connected), then $f \equiv 0$.

$k+1$ with coefficients T. Regarding the tensor T as the parameters of the polynomial system, let $\mathcal{N}(T)$ denote the number of geometrically isolated solutions to the corresponding (2.11) over the algebraically closed complex space. Then it is known by using a continuation argument that $\mathcal{N}(T)$ is the same, say, $\mathcal{N}$, for almost all $T \in \mathbb{C}^{I_1\times\cdots\times I_k}$ and that $\mathcal{N}(T) \leq \mathcal{N}$ for all $T \in \mathbb{C}^{I_1\times\cdots\times I_k}$. Furthermore, the subset of $\mathbb{C}^{I_1\times\cdots\times I_k}$, where $\mathcal{N}(T) = \mathcal{N}$, is a Zariski open set, that is, the exceptional subset of tensors $T \in \mathbb{C}^{I_1\times\cdots\times I_k}$, where $\mathcal{N}(T) < \mathcal{N}$ is an affine algebraic set contained within an algebraic set of codimension one [30, Theorem 7.1.1].

We are interested only in real-valued tensors. Since $\mathbb{R}^n$ is Zariski dense in $\mathbb{C}^n$, the above statements hold for almost all $T \in \mathbb{R}^{I_1\times\cdots\times I_k}$, except that the number of real-valued isolated solutions varies as a function of T and is no longer a constant. For our application, we only need the fact that the real roots of (2.11) are geometrically isolated for generic T.

5.4. An example of exception. Theorem 3.4 holds for almost all tensors T. It might be of interest to demonstrate an exceptional case in this section.

Let T be the tensor given by the tensor product of the identity matrix I_n with itself. This is an order-4 tensor. It is easy to see identities such as

$$\begin{aligned}
\lambda = \left\langle T, \mathbf{u}^{(1)}\otimes\ldots\otimes\mathbf{u}^{(4)}\right\rangle &= \left(\mathbf{u}^{(1)^\top}\mathbf{u}^{(2)}\right)\left(\mathbf{u}^{(3)^\top}\mathbf{u}^{(4)}\right),\\
T\circledast_1\left(\mathbf{u}^{(2)}\otimes\mathbf{u}^{(3)}\otimes\otimes\mathbf{u}^{(4)}\right) &= \left(\mathbf{u}^{(3)^\top}\mathbf{u}^{(4)}\right)\mathbf{u}^{(2)},
\end{aligned}$$

and so on. The corresponding four nonlinear equations (2.11) for stationary points become

$$\begin{aligned}
\left(\mathbf{u}^{(3)^\top}\mathbf{u}^{(4)}\right)\mathbf{u}^{(2)} &= \left(\mathbf{u}^{(1)^\top}\mathbf{u}^{(2)}\right)\left(\mathbf{u}^{(3)^\top}\mathbf{u}^{(4)}\right)\mathbf{u}^{(1)},\\
\left(\mathbf{u}^{(3)^\top}\mathbf{u}^{(4)}\right)\mathbf{u}^{(1)} &= \left(\mathbf{u}^{(1)^\top}\mathbf{u}^{(2)}\right)\left(\mathbf{u}^{(3)^\top}\mathbf{u}^{(4)}\right)\mathbf{u}^{(2)},\\
\left(\mathbf{u}^{(1)^\top}\mathbf{u}^{(2)}\right)\mathbf{u}^{(4)} &= \left(\mathbf{u}^{(1)^\top}\mathbf{u}^{(2)}\right)\left(\mathbf{u}^{(3)^\top}\mathbf{u}^{(4)}\right)\mathbf{u}^{(3)},\\
\left(\mathbf{u}^{(1)^\top}\mathbf{u}^{(2)}\right)\mathbf{u}^{(3)} &= \left(\mathbf{u}^{(1)^\top}\mathbf{u}^{(2)}\right)\left(\mathbf{u}^{(3)^\top}\mathbf{u}^{(4)}\right)\mathbf{u}^{(4)}.
\end{aligned}$$

Assuming $\lambda \neq 0$, it then follows that the set $\mathcal{C}$ of stationary points contains nonisolated solutions $\mathbf{u}^{(1)} = \pm\mathbf{u}^{(2)}$ and $\mathbf{u}^{(3)} = \pm\mathbf{u}^{(4)}$.

Furthermore, the ALS algorithm generates a sequence according to the scheme

$$\mathbf{u}^{(1)}_{[p+1]} = \pm\mathbf{u}^{(2)}_{[p]}, \quad \mathbf{u}^{(2)}_{[p+1]} = \mathbf{u}^{(2)}_{[p]}, \quad \mathbf{u}^{(3)}_{[p+1]} = \pm\mathbf{u}^{(4)}_{[p]}, \quad \mathbf{u}^{(4)}_{[p+1]} = \mathbf{u}^{(4)}_{[p]},$$

where the sign $\pm$ depends on the sign of the initial $\mathbf{u}^{(3)^\top}_{[0]}\mathbf{u}^{(4)}_{[0]}$. Thus, this simple example also demonstrates a situation where, even though the stationary points are not isolated, the ALS method converges in one iteration.

REFERENCES

[1] P.-A. Absil, R. Mahony, and B. Andrews, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM J. Optim., 16 (2005), pp. 531–547.

[2] 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.

[3] R. Chill, *On the Łojasiewicz–Simon gradient inequality*, J. Funct. Anal., 201 (2003), pp. 572–601.

[4] M. T. Chu and K. R. Driessel, *The projected gradient method for least squares matrix approximations with spectral constraints*, SIAM J. Numer. Anal., 27 (1990), pp. 1050–1060.

[5] M. T. Chu and J. L. Watterson, *On a multivariate eigenvalue problem.* I. *Algebraic theory and a power method*, SIAM J. Sci. Comput., 14 (1993), pp. 1089–1106.

[6] P. Comon, X. Luciani, and A. L. F. de Almeida, *Tensor decompositions, alternating least squares and other tales*, J. Chemometrics, 23 (2009), pp. 393–405.

[7] 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.

[8] P. Comon, J. M. F. ten Berge, L. De Lathauwer, and J. Castaing, *Generic and typical ranks of multi-way arrays*, Linear Algebra Appl., 430 (2009), pp. 2997–3007.

[9] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.

[10] L. De Lathauwer, B. De Moor, and J. Vandewalle, *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.

[11] 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.

[12] 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.

[13] L. Grippo and M. Sciandrone, *On the convergence of the block nonlinear Gauss–Seidel method under convex constraints*, Oper. Res. Lett., 26 (2000), pp. 127–136.

[14] V. Guillemin and A. Pollack, *Differential Topology*, AMS Chelsea Publishing, Providence, RI, 2010.

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

[16] R. Hartshorne, *Algebraic Geometry*, Grad. Texts in Math. 52, Springer-Verlag, New York, 1977.

[17] F. L. Hitchcock, *The expression of a tensor or a polyadic as a sum of products*, J. Math. Phys., 6 (1927), pp. 164–189.

[18] H. B. Keller, *Geometrically isolated nonisolated solutions and their approximation*, SIAM J. Numer. Anal., 18 (1981), pp. 822–838.

[19] H. A. L. Kiers, *Towards a standardized notation and terminology in multiway analysis*, J. Chemometrics, 14 (2000), pp. 105–122.

[20] 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.

[21] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

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

[23] J. B. Kruskal, *Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics*, Linear Algebra Appl., 18 (1977), pp. 95–138.

[24] 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 Appl., 269 (1998), pp. 307–329.

[25] L.-H. Lim, *Singular values and eigenvalues of tensors: A variational approach*, in Proceedings of the 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005, pp. 129–132.

[26] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, in Les Équations aux Dérivées Partielles, Éditions du Centre National de la Recherche Scientifique, Paris, 1963, pp. 87–89.

[27] S. Łojasiewicz and M.-A. Zurro, *On the gradient inequality*, Bull. Pol. Acad. Sci. Math., 47 (1999), pp. 143–145.

[28] M. J. Mohlenkamp, *Musings on multilinear fitting*, Linear Algebra Appl., 438 (2013), pp. 834–852.

[29] M. Pierre, *Quelques applications de l'inégalit´e de Lojasiewicz à des discrétisations d'EDP*, in Proceedings of SMAI, 2011; also available online from http://smai.emath.fr/smai2011/slides/mpierre/Slides.pdf.

[30] A. J. Sommese and C. W. Wampler, II, *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*, World Scientific, Hackensack, NJ, 2005.

[31] M. Sørensen, L. D. Lathauwer, P. Comon, S. Icart, and L. Deneire, *Canonical Polyadic Decomposition with a Columnwise Orthonormal Factor Matrix*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 1190–1213.

[32] A. Stegeman and P. Comon, *Subtracting a best rank-*1 *approximation may increase tensor rank*, Linear Algebra Appl., 433 (2010), pp. 1276–1300.

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

[34] A. Uschmajew, *Local convergence of the alternating least squares algorithm for canonical tensor approximation*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 639–652.

[35] T. Zhang and G. H. Golub, *Rank-one approximation to high order tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 534–550.