Electronic Transactions on Numerical Analysis.
Volume 48, pp. 1–14, 2018.

ISSN 1068–9613.
DOI: 10.1553/etna_vol48s1

ETNA
Kent State University and
Johann Radon Institute (RICAM)

ON ACCELERATING THE REGULARIZED ALTERNATING LEAST-SQUARES ALGORITHM FOR TENSORS*

XIAOFEI WANG†, CARMELIZA NAVASCA‡, AND STEFAN KINDERMANN§

Abstract. In this paper, we discuss the acceleration of the regularized alternating least-squares (RALS) algorithm for tensor approximations. We propose a fast iterative method using an Aitken-Stefensen-like update for the regularized algorithm. Through numerical experiments, a faster convergence rate for the accelerated version is demonstrated in comparison to both the standard and regularized alternating least-squares algorithms. In addition, we analyze global convergence based on the Kurdyka-Łojasiewicz inequality, and we show that the RALS algorithm has a linear local convergence rate.

Key words. alternating least-squares, Kurdyka-Łojasiewicz inequality, tensor approximation

AMS subject classifications. 15A69, 65F30

1. Introduction.

Given a third-order tensor $\mathcal{T} \in \mathbb{R}^{I\times J\times K}$, we want to find the best approximation of $\mathcal{T}$ with r rank-one components. This tensor approximation can be posed as an optimization problem:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|\mathcal{T} - \sum_{s=1}^{r} \mathbf{a}_s \circ \mathbf{b}_s \circ \mathbf{c}_s\|_F^2 \\ &\text{subject to } \mathbf{a}_s \in \mathbb{R}^I, \mathbf{b}_s \in \mathbb{R}^J, \mathbf{c}_s \in \mathbb{R}^K, \quad s = 1, \cdots, r, \end{aligned}$$

where $\mathbf{a}_s \circ \mathbf{b}_s \circ \mathbf{c}_s$ is a rank-one tensor generated by taking the outer products of three vectors, $\mathbf{a}_s, \mathbf{b}_s$, and $\mathbf{c}_s$; see below. A global minimizer of the objective functional may not exist due to the ill-posedness [9, 18] of a low-rank approximation, but developing algorithms to detect local minimizers or critical points of the objective functional is important for both theoretical research and practical application of tensor computations [14].

The conventional method for solving this problem, the alternating least-squares (ALS) algorithm [6, 11], which was proposed 45 years ago, remains the workhorse for computing tensor approximations and decompositions. It is based on iteratively solving least-squares subproblems of the original nonlinear objective functional using a Gauss-Seidel updating scheme. The subproblems are obtained through matricizing the given tensor and the rank-one tensor components. Under an assumption on the Hessian of the objective functional, it was shown in [26] that the ALS algorithm has a linear local convergence rate. Despite the success of the ALS algorithm, it has some shortcomings [8, 24]. The non-uniqueness of the solution within the inner iterations of the ALS can substantially slow down convergence. This non-uniqueness can be avoided by introducing a Tikhonov-regularized term to the objective functional [18, 24]. However, this new update mechanism with such a term included cannot guarantee that the local minimizer is also a fixed-point of the ALS update operator. Another regularization [20, 16] was proposed to handle the ALS algorithm by introducing a proximal

*Received July 21, 2017. Accepted December 11, 2017. Published online on February 16, 2018. Recommended by L. Reichel. This work was supported by National Natural Science Foundation of China (Grants No. 11401092), China Scholarship Council (Grants No.201406625025)

†Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Renmin Street 5268, Changchun, China (wangxf341@nenu.edu.cn).

‡Department of Mathematics, University of Alabama at Birmingham, 1300 University Boulevard, Birmingham, AL, USA (cnavasca@uab.edu).

§Industrial Mathematics Institute, Johannes Kepler Universitat Linz, Altenbergerstrasse 69, A-4040 Linz, Austria (kindermann@indmath.uni-linz.ac.at).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

term into every subproblem instead of directly into the objective functional. This regularized version of the ALS algorithm is called the regularized alternating least-squares (RALS) algorithm. It was shown in [16] that any limit point of every convergent subsequence from the RALS algorithm is a critical point of the objective functional.

Both the ALS and RALS algorithms update one block of variables at each iteration while fixing other blocks. Thus, these two algorithms can be considered within the framework of several alternating block minimization techniques [2, 3, 27]. The Kurdyka-Łojasiewicz (KL) inequality [19] is the essential tool to show global convergence of the ALS method. Attouch et al. [2, 3] study the convergence properties of alternating proximal minimization algorithms for nonconvex structured functions. In [27], Xu and Yin develop the block coordinate descent method with the Gauss-Seidel updating sweep for block multi-convex functions with applications to nonnegative tensor factorization and tensor completion. Instead of updating all the blocks in each loop as in [2, 3, 27], an alternative approach is the maximum block improvement (MBI) method [7], which only updates *the maximally improving* block per loop. In [17], MBI was shown to handle tensor optimization models with spherical constraints. Under some mild assumptions, Li et al. [17] show that MBI achieves global convergence and a linear local convergent rate. Here we consider the convergence properties of the regularized alternating least-squares (RALS) method in case the regularization parameter is static. We show global convergence of the RALS algorithm within the framework of proximal alternating minimization [2, 3]. The rate of this global convergence depends on the exponent in the Kurdyka-Łojasiewicz inequality. We prove that the global convergence rate is either linear or sublinear, but to further discern between these cases relies on a priori knowledge of the exponent of the KL inequality. In the appendix, we discuss the local convergence theory of RALS, namely, we prove that if the sequence is close enough to a local minimizer, then the RALS algorithm has a linear local convergence rate.

Moreover, in this paper, we propose a new acceleration version of RALS by extending the Aitken-Stefensen acceleration formula to matrix form. The corresponding numerical simulation results illustrate the effectiveness of our acceleration method. In addition, the new fast method outperforms Nesterov-accelerated [21] ALS and RALS.

This paper is organized as follows. In Section 2, we introduce some notations and terminologies for the RALS algorithm for tensor approximation. In Section 3 we propose an accelerated version of the algorithm. A simulation experiment is presented in Section 4. In Section 5, we discuss global convergence rates of the algorithm. Finally, in Section 6 we summarize our conclusions and indicate some remaining problems.

2. The RALS algorithm for tensor approximation. We focus on third-order tensors $\mathcal{T} = (t_{ijk}) \in \mathbb{R}^{I\times J\times K}$ with three indices $1 \le i \le I, 1 \le j \le J$, and $1 \le k \le K$, but all the methods proposed here can be applied to tensors of arbitrary d-th order. A third-order tensor $\mathcal{T}$ has column, row, and tube fibers, which are defined by fixing every index but one and denoted by $\mathbf{t}_{:jk}$, $\mathbf{t}_{i:k}$, and $\mathbf{t}_{ij:}$, respectively. Correspondingly, we obtain three matricizations of $\mathcal{T}$:

$$\begin{aligned}
\mathbf{T}_{(1)} &= [\mathbf{t}_{:11}, \cdots, \mathbf{t}_{:J1}, \mathbf{t}_{:12}, \cdots, \mathbf{t}_{:J2}, \cdots, \mathbf{t}_{:1K}, \cdots, \mathbf{t}_{:JK}], && \mathbf{T}_{(1)} \in \mathbb{R}^{I\times JK},\\
\mathbf{T}_{(2)} &= [\mathbf{t}_{1:1}, \cdots, \mathbf{t}_{I:1}, \mathbf{t}_{1:2}, \cdots, \mathbf{t}_{I:2}, \cdots, \mathbf{t}_{1:K}, \cdots, \mathbf{t}_{I:K}], && \mathbf{T}_{(2)} \in \mathbb{R}^{J\times IK},\\
\mathbf{T}_{(3)} &= [\mathbf{t}_{11:}, \cdots, \mathbf{t}_{I1:}, \mathbf{t}_{12:}, \cdots, \mathbf{t}_{I2:}, \cdots, \mathbf{t}_{1J:}, \cdots, \mathbf{t}_{IJ:}], && \mathbf{T}_{(3)} \in \mathbb{R}^{K\times IJ}.
\end{aligned}$$

The outer product $\mathbf{a}\circ\mathbf{b}\circ\mathbf{c} \in \mathbb{R}^{I\times J\times K}$ of three nonzero vectors $\mathbf{a} \in \mathbb{R}^I, \mathbf{b} \in \mathbb{R}^J$, and $\mathbf{c} \in \mathbb{R}^K$ is called a rank-one tensor and is defined by $(\mathbf{a}\circ\mathbf{b}\circ\mathbf{c})_{i,j,k} = a_i b_j c_k$ for all the indices i, j, k in their corresponding index ranges. A canonical polyadic (CP) decomposition

ETNA
Kent State University and
Johann Radon Institute (RICAM)

of $\mathcal{T} \in \mathbb{R}^{I\times J\times K}$ expresses $\mathcal{T}$ as a sum of rank-one outer products:

$$\mathcal{T} = \sum_{s=1}^{r} \mathbf{a}_s \circ \mathbf{b}_s \circ \mathbf{c}_s, \tag{2.1}$$

where $\mathbf{a}_s \in \mathbb{R}^I, \mathbf{b}_s \in \mathbb{R}^J, \mathbf{c}_s \in \mathbb{R}^K$, for $1 \le s \le r$. Every outer product $\mathbf{a}_s \circ \mathbf{b}_s \circ \mathbf{c}_s$ is a rank-one component. The positive integer r is the number of rank-one tensor components of $\mathcal{T}$.

The Khatri-Rao product of two matrices $\mathbf{A} \in \mathbb{R}^{I\times r}$ and $\mathbf{B} \in \mathbb{R}^{J\times r}$ is defined as

$$\mathbf{A} \odot \mathbf{B} = (\mathbf{a}_1 \otimes \mathbf{b}_1, \cdots, \mathbf{a}_R \otimes \mathbf{b}_R) \in \mathbb{R}^{IJ\times r},$$

where the symbol "$\otimes$" denotes the Kronecker product:

$$\mathbf{a} \otimes \mathbf{b} = (a_1 b_1, \cdots, a_1 b_J, \cdots, a_I b_1, \cdots, a_I b_J)^T.$$

Using this Khatri-Rao product, the CP decomposition (2.1) can be equivalently expressed by one of the three matricization forms of the tensor $\mathcal{T}$:

$$\mathbf{T}_{(1)} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T, \qquad \mathbf{T}_{(2)} = \mathbf{B}(\mathbf{C} \odot \mathbf{A})^T, \qquad \mathbf{T}_{(3)} = \mathbf{C}(\mathbf{B} \odot \mathbf{A})^T,$$

where $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_r) \in \mathbb{R}^{I\times r}$, $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_r) \in \mathbb{R}^{J\times r}$, and $\mathbf{C} = (\mathbf{c}_1, \cdots, \mathbf{c}_r) \in \mathbb{R}^{K\times r}$ are called the factor matrices of the tensor $\mathcal{T}$.

Let $\mathscr{X} = \mathbb{R}^{I\times r} \times \mathbb{R}^{J\times r} \times \mathbb{R}^{K\times r}$, where r is any given positive integer, and let the elements of $\mathscr{X}$ be denoted by $\mathbf{x} = (\mathbf{A}, \mathbf{B}, \mathbf{C})$, where $\mathbf{A} \in \mathbb{R}^{I\times r}, \mathbf{B} \in \mathbb{R}^{J\times r}, \mathbf{C} \in \mathbb{R}^{K\times r}$. Note that $\mathbf{x}$ can also be viewed as a vector in $\mathbb{R}^{r(I+J+K)}$. Given a tensor $\mathcal{T} \in \mathbb{R}^{I\times J\times K}$, we consider its approximation by using the sum of r rank-one components $\sum_{s=1}^{r} \mathbf{a}_s \circ \mathbf{b}_s \circ \mathbf{c}_s$, and define a residual functional $f : \mathscr{X} \to \mathbb{R}$ by

$$f(\mathbf{x}) = f(\mathbf{A}, \mathbf{B}, \mathbf{C}) \to \frac{1}{2}\|\mathcal{T} - \sum_{s=1}^{r} \mathbf{a}_s \circ \mathbf{b}_s \circ \mathbf{c}_s\|_F^2,$$

where the vectors $\mathbf{a}_s, \mathbf{b}_s, \mathbf{c}_s$ are columns of $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, respectively, and $\|\cdot\|_F$ is the tensor Frobenius norm. There may exist a local minimizer $\mathbf{x}^* = (\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ of $f(\mathbf{A}, \mathbf{B}, \mathbf{C})$, which is hence also a critical point of $f(\mathbf{x})$ such that $\nabla f(\mathbf{x}^*) = 0$ since f is a polynomial function. Denote $\sum_{s=1}^{r} \mathbf{a}_s^* \otimes \mathbf{b}_s^* \otimes \mathbf{c}_s^*$ an optimal approximation of the tensor $\mathcal{T}$ with rank at most r, where the vectors $\mathbf{a}_s^*, \mathbf{b}_s^*, \mathbf{c}_s^*$ are columns of some matrices $\mathbf{A}^*, \mathbf{B}^*$, and $\mathbf{C}^*$, respectively.

The approximation of a given tensor is implemented by the alternating least-squares (ALS) algorithm. Given a starting point $\mathbf{x}^{(0)} = (\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{C}^{(0)})$, we solve three subproblems iteratively:

(2.2)

$$\mathbf{A}^{(n+1)} = \arg\min_{\mathbf{A}\in\mathbb{R}^{I\times r}} f(\mathbf{A}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)}) = \arg\min_{\mathbf{A}\in\mathbb{R}^{I\times r}} \frac{1}{2}\|\mathbf{T}_{(1)} - \mathbf{A}(\mathbf{C}^{(n)} \odot \mathbf{B}^{(n)^T})\|_F^2,$$

$$\mathbf{B}^{(n+1)} = \arg\min_{\mathbf{B}\in\mathbb{R}^{J\times r}} f(\mathbf{A}^{(n+1)}, \mathbf{B}, \mathbf{C}^{(n)}) = \arg\min_{\mathbf{B}\in\mathbb{R}^{J\times r}} \frac{1}{2}\|\mathbf{T}_{(2)} - \mathbf{B}(\mathbf{C}^{(n)} \odot \mathbf{A}^{(n+1)^T})\|_F^2,$$

$$\mathbf{C}^{(n+1)} = \arg\min_{\mathbf{C}\in\mathbb{R}^{K\times r}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}) = \arg\min_{\mathbf{C}\in\mathbb{R}^{K\times r}} \frac{1}{2}\|\mathbf{T}_{(3)} - \mathbf{C}(\mathbf{B}^{(n+1)} \odot \mathbf{A}^{(n+1)^T})\|_F^2.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

If every optimization problem possesses a unique solution, then one loop of (2.2) defines an operator $S_{ALS}(\cdot)$ [26] via

$$(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}^{(n+1)}) = \mathbf{x}^{(n+1)} = S_{ALS}(\mathbf{x}^{(n)}) = S_{ALS}(\mathbf{A}^{(n)}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)}), \tag{2.3}$$

where the three matrices

$$\begin{aligned}
\mathbf{A}^{(n+1)} &= \left(\mathbf{T}_{(1)}(\mathbf{C}^{(n)} \odot \mathbf{B}^{(n)})\right)\left((\mathbf{C}^{(n)} \odot \mathbf{B}^{(n)})^T(\mathbf{C}^{(n)} \odot \mathbf{B}^{(n)})\right)^{-1},\\
\mathbf{B}^{(n+1)} &= \left(\mathbf{T}_{(2)}(\mathbf{C}^{(n)} \odot \mathbf{A}^{(n+1)})\right)\left((\mathbf{C}^{(n)} \odot \mathbf{A}^{(n+1)})^T(\mathbf{C}^{(n)} \odot \mathbf{A}^{(n+1)})\right)^{-1},\\
\mathbf{C}^{(n+1)} &= \left(\mathbf{T}_{(3)}(\mathbf{B}^{(n+1)} \odot \mathbf{A}^{(n+1)})\right)\left((\mathbf{B}^{(n+1)} \odot \mathbf{A}^{(n+1)})^T(\mathbf{B}^{(n+1)} \odot \mathbf{A}^{(n+1)})\right)^{-1}
\end{aligned} \tag{2.4}$$

are the least-squares solutions of (2.2). Note that the inversion in (2.4) may not exist due to collinearity of the columns in the factor matrices, thus, we consider the generalized Moore-Penrose inverse in this case.

Since the computations in steps (2.2) may not give a unique solution, an extra regularized term [3, 16, 20] is added in every step to eliminate such a non-uniqueness. This regularized ALS algorithm (RALS) is defined as follows:

$$\begin{aligned}
\mathbf{A}^{(n+1)} &= \underset{\mathbf{A}\in\mathbb{R}^{I\times r}}{\arg\min}\, f(\mathbf{A}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)}) + \frac{1}{2}\lambda\|\mathbf{A} - \mathbf{A}^{(n)}\|_F^2,\\
\mathbf{B}^{(n+1)} &= \underset{\mathbf{B}\in\mathbb{R}^{J\times r}}{\arg\min}\, f(\mathbf{A}^{(n+1)}, \mathbf{B}, \mathbf{C}^{(n)}) + \frac{1}{2}\lambda\|\mathbf{B} - \mathbf{B}^{(n)}\|_F^2,\\
\mathbf{C}^{(n+1)} &= \underset{\mathbf{C}\in\mathbb{R}^{K\times r}}{\arg\min}\, f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}) + \frac{1}{2}\lambda\|\mathbf{C} - \mathbf{C}^{(n)}\|_F^2,
\end{aligned} \tag{2.5}$$

where $\lambda > 0$ is a regularization parameter. Our work is based on this RALS model and addresses the case when the regularization parameter λ is static. It is easy to verify that every subproblem in (2.5) must have a unique solution because of strict convexity. We express the update of (2.5) for $\mathbf{A}, \mathbf{B}, \mathbf{C}$ by using the operator $S(\cdot)$:

$$(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}^{(n+1)}) = \mathbf{x}^{(n+1)} = S(\mathbf{x}^{(n)}) = S(\mathbf{A}^{(n)}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)}), \tag{2.6}$$

where the three matrices

$$\begin{aligned}
\mathbf{A}^{(n+1)} &= \left(\mathbf{T}_{(1)}(\mathbf{C}^{(n)} \odot \mathbf{B}^{(n)}) + \lambda\mathbf{A}^{(n)}\right)\left((\mathbf{C}^{(n)} \odot \mathbf{B}^{(n)})^T(\mathbf{C}^{(n)} \odot \mathbf{B}^{(n)}) + \lambda\mathbf{I}\right)^{-1},\\
\mathbf{B}^{(n+1)} &= \left(\mathbf{T}_{(2)}(\mathbf{C}^{(n)} \odot \mathbf{A}^{(n+1)}) + \lambda\mathbf{B}^{(n)}\right)\left((\mathbf{C}^{(n)} \odot \mathbf{A}^{(n+1)})^T(\mathbf{C}^{(n)} \odot \mathbf{A}^{(n+1)}) + \lambda\mathbf{I}\right)^{-1},\\
\mathbf{C}^{(n+1)} &= \left(\mathbf{T}_{(3)}(\mathbf{B}^{(n+1)} \odot \mathbf{A}^{(n+1)}) + \lambda\mathbf{C}^{(n)}\right)\left((\mathbf{B}^{(n+1)} \odot \mathbf{A}^{(n+1)})^T(\mathbf{B}^{(n+1)} \odot \mathbf{A}^{(n+1)}) + \lambda\mathbf{I}\right)^{-1}
\end{aligned} \tag{2.7}$$

are the least-squares solutions of (2.5).

The RALS algorithm can be viewed as a proximal regularization of a three-block Gauss-Seidel method for minimizing $f(\mathbf{A}, \mathbf{B}, \mathbf{C})$. In Section 5, we show global convergence of the RALS algorithm within the framework of proximal alternating minimization [2, 5].

3. Acceleration of the RALS algorithm. In this section, we suggest an acceleration technique for the RALS algorithm. Our acceleration method is loosely based on the Aitken-Stefensen formula [13], which is a conventional acceleration technique for numerical computations. In the scalar case, for a given convergent sequence $\{x^{(n)}\}_{n\in\mathbb{N}}$, a new sequence $\{y^{(n)}\}_{n\in\mathbb{N}}$ is generated by

$$y^{(n)} = x^{(n)} - \frac{(\triangle x^{(n)})^2}{\triangle^2 x^{(n)}}, \tag{3.1}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

where $\triangle x^{(n)} = x^{(n+1)} - x^{(n)}$ and $\triangle^2 x^{(n)} = x^{(n+2)} - 2x^{(n+1)} + x^{(n)}$. For fixed-point iterations, the Aitken-Steffensen acceleration (3.1) can achieve a quadratic convergent rate [13] without the use of derivative terms.

The generalization of the Aitken-Stefensen process to a k-dimensional sequence requires the following iterative formula:

$$\mathbf{y}^{(n)} = \mathbf{x}^{(n)} - \triangle\mathbf{X}^{(n)}(\triangle^2\mathbf{X}^{(n)})^{-1}\triangle\mathbf{x}^{(n)}, \tag{3.2}$$

where

$$\begin{aligned}
\triangle\mathbf{x}^{(n)} &= \mathbf{x}^{(n+1)} - \mathbf{x}^{(n)},\\
\triangle\mathbf{X}^{(n)} &= (\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}, \cdots, \mathbf{x}^{(n+k)} - \mathbf{x}^{(n+k-1)}), \qquad \text{and}\\
\triangle^2\mathbf{X}^{(n)} &= (\mathbf{x}^{(n+2)} - 2\mathbf{x}^{(n+1)} + \mathbf{x}^{n}, \cdots, \mathbf{x}^{(n+k+1)} - 2\mathbf{x}^{(n+k)} + \mathbf{x}^{(n+k-1)}).
\end{aligned}$$

The formula (3.2) for $\{\mathbf{y}^{(n)}\}_{n\in\mathbb{N}}$ also has a quadratic convergence rate under five basic assumptions [22]. Although the Aitken-Stefensen process for k-dimensional sequences theoretically has a fast convergent rate, it has two main drawbacks in the practical implementation. One is that in order to compute $\mathbf{y}^{(n)}$, an a priori set of sequences is needed, namely, $\mathbf{x}^{(1)}$ to $\mathbf{x}^{(n+k+1)}$. In case the dimension k of the vectorspace is large, the practical implementation will be time-consuming especially when facing a complicated updating map. The other is that this iterative process may be invalid if the original sequence $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$ converges fast and the dimension k is large enough such that $\mathbf{x}^{(n+k+1)} - 2\mathbf{x}^{(n+k)} + \mathbf{x}^{(n+k-1)}$ is close to zero and $\triangle^2\mathbf{X}^{(n)}$ is (almost) singular. So although the Aitken-Stefensen method can be directly applied to the acceleration of the $r(I+J+K)$-dimensional sequence $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$ generated by the RALS algorithm, it does not work well, especially when I, J, K, r are large. For example, if $I = J = K = 20$ and $r = 10$, then the dimension of k is 600. To compute the initial vector of $\mathbf{y}^{(0)}$ from $\mathbf{x}^{(0)}$, we need to know 601 vectors from $\mathbf{x}^{(1)}$ to $\mathbf{x}^{(601)}$. But the original sequence $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$ from RALS may have already converged before $n = 601$.

To obviate these drawbacks of the recursive formula (3.2) of vectors, we utilize the matrix format of the update (2.7) for the RALS algorithm and propose a matrix-based Aitken-Stefensen acceleration formula. We denote the $(I+J+K)\times r$ matrix $(\mathbf{A}^{(n)^T}, \mathbf{B}^{(n)^T}, \mathbf{C}^{(n)^T})^T$ by $\mathbf{X}^{(n)}$, and define the update by

$$\mathbf{X}_*^{(n+1)} = \mathbf{X}^{(n)} - \mathbf{Z}^{(n)}, \tag{3.3}$$

where $\mathbf{Z}^{(n)}$ is a solution of the linear system

$$\mathbf{Z}^{(n)}\left(S(S(\mathbf{X}^{(n)})) - 2S(\mathbf{X}^{(n)}) + \mathbf{X}^{(n)}\right)^T = (S(\mathbf{X}^{(n)}) - \mathbf{X}^{(n)})(S(\mathbf{X}^{(n)}) - \mathbf{X}^{(n)})^T. \tag{3.4}$$

Here the matrix $\mathbf{Z}^{(n)}$ can be understood as a small perturbation of $\mathbf{X}^{(n)}$ towards $\mathbf{X}_*^{(n+1)}$ since $\|S(\mathbf{X}^{(n)}) - \mathbf{X}^{(n)}\|_F^2$ is small when $\mathbf{X}^{(n)}$ is close to a fixed-point of S (as defined by (2.6)). Note that $S(\mathbf{X}^{(n)})$ is based on RALS, and we express the new update (3.3) from $\mathbf{X}^{(n)}$ to $\mathbf{X}_*^{(n+1)}$ by an operator T:

$$\mathbf{X}_*^{(n+1)} = T(\mathbf{X}^{(n)}).$$

It can be verified that a fixed-point of the operator T is also a fixed-point of the operator S.

Notice that besides one extra update from $S(\mathbf{X}^{(n)})$ to $(S(S(\mathbf{X}^{(n)}))$, the formula (3.3) involves solving a large linear system (3.4) with the matrix $(S(S(\mathbf{X}^{(n)})) - 2S(\mathbf{X}^{(n)}) + \mathbf{X}^{(n)})^T$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

of size $r \times (I + J + K)$. If (3.3) is computed in each step of the algorithm, then the whole computational cost of the practical implementation will be huge. So in the following Algorithm 1, we employ the update formula (3.3) not at every step n, but only after a fixed (specified) number of iterations and if the residual is small enough. From another perspective, the formula allows the outer iteration of the (R)ALS algorithm to *jump out* from the linear convergent regions. The residual gap of these perturbations are quickly eliminated. Several numerical experiments are presented in the next section.

Algorithm 1 Acceleration of RALS (RALS-A)

Input: A third order tensor $\mathcal{T} \in \mathbb{R}^{I\times J\times K}$, the number r of rank-one components, an interval positive integer q, and a upper bound $\alpha \in \mathbb{R}$;

Output: Three matrices $\mathbf{A} \in \mathbb{R}^{I\times r}, \mathbf{B} \in \mathbb{R}^{J\times r}, \mathbf{C} \in \mathbb{R}^{K\times r}$;

Give initial matrices $(\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{C}^{(0)})$ and let $\mathbf{X}^{(0)} = (\mathbf{A}^{(0)^T}, \mathbf{B}^{(0)^T}, \mathbf{C}^{(0)^T})^T$ and set the error square as err $= \alpha$.
Update step:
for $n = 1, \cdots$ **do**
  if err $< \alpha$ and n mod $q = 0$ **do**
    Compute the matrices $S(S(\mathbf{X}^{(n)}))$ and $S(\mathbf{X}^{(n)})$ from $\mathbf{X}^{(n)}$.
    Compute the matrix $\mathbf{X}_*^{(n+1)}$ by using (3.3).
    $\mathbf{X}^{(n+1)} = \mathbf{X}_*^{(n+1)}$.
  else do
    Compute the matrix $S(\mathbf{X}^{(n)})$ from $\mathbf{X}^{(n)}$.
    $\mathbf{X}^{(n+1)} = S(\mathbf{X}^{(n)})$.
  end if
  err $= \|\mathbf{X}^{(n+1)} - \mathbf{X}^{(n)}\|_F^2$.
end for
$\mathbf{A} = \mathbf{A}^{(n)}$, $\mathbf{B} = \mathbf{B}^{(n)}$, $\mathbf{C} = \mathbf{C}^{(n)}$.
return Three matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$.

4. Numerical experiments. In this section we demonstrate the simulation experiments of the ALS, RALS algorithms and their accelerated versions. The experiments are done with Matlab and implemented on a desktop computer with an Intel i5 CPU 3.3GHz CPU and 8G memory. All of these algorithms use a tolerance error of 1×10^{-12} as a stopping criterion for the update

$$\|\mathbf{X}^{(n)} - \mathbf{X}^{(n-1)}\|_F^2 = \|\mathbf{A}^{(n)} - \mathbf{A}^{(n-1)}\|_F^2 + \|\mathbf{B}^{(n)} - \mathbf{B}^{(n-1)}\|_F^2 + \|\mathbf{C}^{(n)} - \mathbf{C}^{(n-1)}\|_F^2$$

between two subsequent iterates. One may use a relative error-based stopping criterion [27] to address the issue of a scale dependence of the input tensor. Algorithm 1 is an accelerated version of the RALS algorithm, and we denote it RALS-A. We can similarly obtain an acceleration of the ALS algorithm called ALS-A. More specifically, ALS-A is realized by replacing the update operator S in Algorithm 1 by the operator S_{ALS} in (2.3). The upper bound α is an input parameter for judging whether the original sequence is already in a linear convergent region. As long as err $< \alpha$, we employ the acceleration update after a fixed number q of iterations. In the simulation experiments, we choose $\alpha = 1 \times 10^{-6}$ and $q = 100$. Besides

ETNA
Kent State University and
Johann Radon Institute (RICAM)

TABLE 4.1
Time costs of ALS, ALS-A, RALS,RALS-A, RALS-L, and RALS-AL.

Algorithm	ALS	ALS-Nes	ALS-A	RALS	RALS-Nes	RALS-A	RALS-L	RALS-AL
$I = 10$	0.59	2.41	0.38	0.89	1.77	0.51	0.59	**0.36**
$I = 20$	0.47	1.20	0.33	0.55	1.17	0.37	0.50	**0.31**
$I = 50$	2.31	7.25	**1.64**	2.57	6.73	1.86	2.55	1.86

our acceleration method, we also consider the Nesterov-type acceleration (RALS-Nes):

$$\begin{aligned}\mathbf{x}^{(n+1)} &= S(\mathbf{x}_*^{(n)}),\\ \mathbf{x}_*^{(n+1)} &= (1-\gamma_n)\mathbf{x}^{(n+1)} + \gamma_n \mathbf{x}^{(n)}\end{aligned}$$

where $\gamma_n = \frac{1-\mu_n}{\mu_{n+1}}, \mu_n = \frac{1+\sqrt{1+4\mu_{n-1}^2}}{2}, \mu_0 = 0$. We can similarly obtain a Nesterov-type acceleration of the ALS algorithm (ALS-Nes) by replacing the update operator S by S_{ALS}.

At first, we consider the time costs of the ALS, ALS-A, ALS-Nes, RALS, RALS-L, RALS-A, and RALS-AL algorithms, where RALS-L and RALS-AL denotes two modified versions of RALS and RALS-A with a monotonically decreasing regularization parameter λ that converges to zero as the iteration number $n \to \infty$. The rank-one component number r is set to 10 and the dimensions $I = J = K$. For each $I = 10, 20, 50$, we perform 100 numerical experiments for these seven algorithms and record the corresponding seven medians of the time costs in seconds. As shown in Table 4.1, the accelerated versions ALS-A and RALS-A perform much better than the original ALS and RALS algorithms. The RALS-L method with decreasing λ has a higher speed than RALS, and RALS-AL is the fasted among all the algorithms based on RALS. The ALS-Nes scheme requires more times than other algorithms. The reason may lie in the fact that the Nesterov-type acceleration is designed for convex optimization [4, 21]. The main objective functional of the RALS is, however, a nonconvex function while only the subproblems are convex.

Secondly, we consider the convergence of the ALS, ALS-A, RALS, and RALS-A algorithms. Two experiments are presented in Figure 4.1 according to the appearance of swamps in the ALS method or not. In each experiment, we set $I = J = K = r = 10$, and all of those algorithms use the same tensor $\mathcal{T} \in \mathbb{R}^{10\times 10\times 10}$ with same initial factor matrices. For the RALS and RALS-A algorithms, the regularization parameter λ is fixed to 1. The plots in Figure 4.1 display the squared error $\|\mathbf{X}^{(n)} - \mathbf{X}^{(n-1)}\|_F^2$ versus the number of iterations n. As one can see, the convergence of the RALS algorithm is linear (see Appendix A), and the acceleration version RALS-A has a higher convergent rate than RALS. This situation is similar for the ALS and ALS-A algorithms. Notice that ALS without swamps performs much better than RALS with a fixed λ. But as demonstrated in the experiments, the RALS algorithm with a decreasing λ has the highest speed; see Table 4.1.

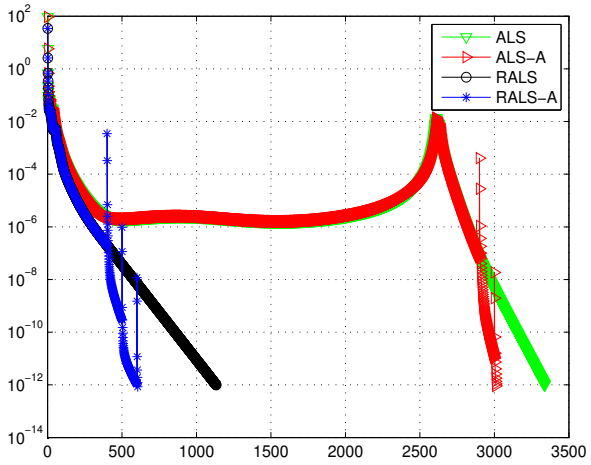
5. Global convergence of RALS. To discuss global convergence of the RALS algorithm, we need the Kurdyka-Łojasiewicz inequality for real-analytic functions. As shown in [19], we have the following proposition for the gradient inequality.

PROPOSITION 5.1 (The Kurdyka-Łojasiewicz inequality). *Let $f(\mathbf{x})$ be a real-analytic function in a neighborhood of $\mathbf{0} \in \mathbb{R}^n$ such that $f(\mathbf{0}) = 0$. Then the following inequality holds for some $0 < \theta < 1$*

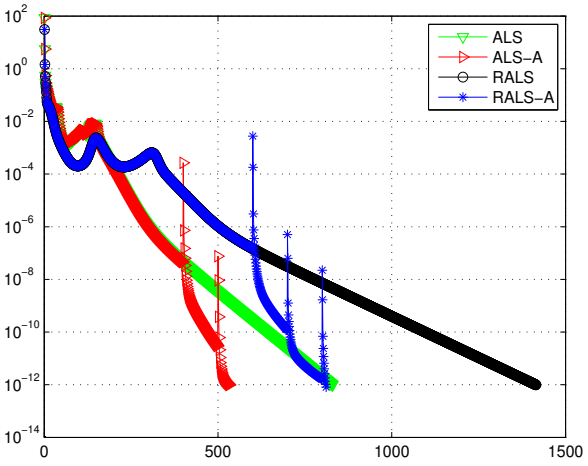
$$|f(\mathbf{x})|^\theta \le \|\nabla f(\mathbf{x})\|$$

in a neighborhood of $\mathbf{0}$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)



(a) ALS with swamp.



(b) ALS without swamp.

FIG. 4.1. *A comparison among ALS, ALS-A, RALS, and RALS-A.*

Clearly, if f is a real-analytic function in a neighborhood of $\mathbf{a} \in \mathbb{R}^n$, then by shifting the origin, we obtain that $g(\mathbf{x}) = f(\mathbf{a}+\mathbf{x}) - f(\mathbf{a})$ is a real-analytic function in a neighborhood of $\mathbf{0} \in \mathbb{R}^n$ and $g(\mathbf{0}) = 0$. From this Proposition 5.1, we find that

$$|f(\mathbf{a}+\mathbf{x}) - f(\mathbf{a})|^{\theta} \leq \|\nabla f(\mathbf{a}+\mathbf{x})\|$$

for any $\mathbf{x}$ in a neighborhood of $\mathbf{0}$. It also follows that $|f(\mathbf{x}) - f(\mathbf{a})|^{\theta} \leq \|\nabla f(\mathbf{x})\|$ for any $\mathbf{x}$ in a neighborhood of $\mathbf{a}$. Thus, we find an equivalent formulation:

PROPOSITION 5.2. *Let $f(\mathbf{x})$ be a real-analytic function on $\mathbb{R}^n$. For any $\mathbf{a} \in \mathbb{R}^n$, there exists a real number $0 < \theta < 1$ and a neighborhood U of $\mathbf{a}$ such that*

$$|f(\mathbf{x}) - f(\mathbf{a})|^{\theta} \leq \|\nabla f(\mathbf{x})\|$$

for any $\mathbf{x} \in U$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

By using Proposition 5.2 and the finite subcover property of compact sets, we arrive at the following statement [5, 10].

PROPOSITION 5.3. *Let E be the set of critical points of f and Γ be a compact and connected subset of E. If f is a real-analytic function on $\mathbb{R}^n$ and $\mathbf{a} \in \Gamma$, then*

1. *for any $\mathbf{b} \in \Gamma$, $f(\mathbf{b}) = f(\mathbf{a}) \triangleq \underline{f}$, and*
2. *there is a neighborhood U of Γ and a real number $0 < \theta < 1$ such that*

$$\forall \mathbf{x} \in U: \quad |f(\mathbf{x}) - \underline{f}|^\theta \leq \|\nabla f(\mathbf{x})\|.$$

In the RALS algorithm, the residual functional $f(\mathbf{x}) = f(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is a polynomial function on $\mathscr{X} = \mathbb{R}^{I\times r} \times \mathbb{R}^{J\times r} \times \mathbb{R}^{K\times r}$. So it is also a real-analytic function on $\mathscr{X}$. Unlike in the work of Li et al. [16], where it is shown that every limit point is a critical point, the following theorem yields global convergence of the RALS algorithm. Its proof is based on the Kurdyka-Łojasiewicz inequality and the proximal alternating minimum technique [2, 3, 5].

THEOREM 5.4. *Let $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$ be the sequence generated by the RALS algorithm. If the sequence $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$ is bounded, then this sequence converges to a critical point $\mathbf{x}^*$ of $f(\mathbf{x})$.*

Proof. In the RALS algorithm, the residual functional

$$f(\mathbf{x}) = f(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \frac{1}{2}\|\mathcal{T} - \sum_{s=1}^{r} \mathbf{a}_s \circ \mathbf{b}_s \circ \mathbf{c}_s\|^2$$

is a polynomial function on $\mathscr{X} = \mathbb{R}^{I\times r} \times \mathbb{R}^{J\times r} \times \mathbb{R}^{K\times r}$, where $\mathbf{a}_s, \mathbf{b}_s, \mathbf{c}_s$ are columns of some matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, respectively. From (2.5), we know that

$$f(\mathbf{x}^{(n)}) - f(\mathbf{x}^{(n+1)}) \geq \frac{1}{2}\lambda\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\|^2 \tag{5.1}$$

and

$$\begin{aligned} \nabla_{\mathbf{A}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)}) + \lambda(\mathbf{A}^{(n+1)} - \mathbf{A}^{(n)}) = 0, \\ \nabla_{\mathbf{B}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}^{(n)}) + \lambda(\mathbf{B}^{(n+1)} - \mathbf{B}^{(n)}) = 0, \\ \nabla_{\mathbf{C}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}^{(n+1)}) + \lambda(\mathbf{C}^{(n+1)} - \mathbf{C}^{(n)}) = 0. \end{aligned} \tag{5.2}$$

From (5.1), we have that $\lim_{n\to\infty} \|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| = 0$ and $\{f(\mathbf{x}^{(n)})\}_{n\in\mathbb{N}}$ is a monotonically decreasing sequence. Let $\underline{f} = \lim_{n\to\infty} f(\mathbf{x}^{(n)})$.

Due to the boundedness of $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$, the first equality in (5.2), and the differentiability of $f(\mathbf{x})$, there exist constants $\lambda_1, \lambda_2 > 0$ and $\mu_1 > 0$ such that

$$\begin{aligned} &\|\nabla_{\mathbf{A}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}^{(n+1)})\|_F \\ &\quad \leq \|\nabla_{\mathbf{A}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}^{(n+1)}) - \nabla_{\mathbf{A}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)})\|_F \\ &\qquad + \|\nabla_{\mathbf{A}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)})\|_F \\ &\quad \leq \lambda_1\|\mathbf{B}^{(n+1)} - \mathbf{B}^{(n)}\|_F + \lambda_2\|\mathbf{C}^{(n+1)} - \mathbf{C}^{(n)}\|_F + \lambda\|\mathbf{A}^{(n+1)} - \mathbf{A}^{(n)}\|_F \\ &\quad \leq \mu_1\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| \end{aligned}$$

for any $n \in \mathbb{N}$. Similarly, there exist constants $\mu_2, \mu_3 > 0$ such that

$$\begin{aligned} \|\nabla_{\mathbf{B}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}^{(n+1)})\|_F \leq \mu_2\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| \\ \|\nabla_{\mathbf{C}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}^{(n+1)})\|_F \leq \mu_3\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\|. \end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

It follows that there exists a constant $d > 0$ such that

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}^{(n+1)})\| \leq d\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| \tag{5.3}$$

for any $n \in \mathbb{N}$.

Denote the limit point set of $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$ by L. From the inequality (5.3), any point in L is a critical point of f. It can also be verified that L is a compact and connected set since $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$ is bounded and $\lim_{n\to\infty} \|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| = 0$. So, from Proposition 5.3, we have $f(\mathbf{x}) = \underline{f}$ for any $\mathbf{x} \in L$, and there is a neighborhood U of L and a real number $0 < \theta < 1$ such that $|f(\mathbf{x}) - \underline{f}|^{\theta} \leq \|\nabla f(\mathbf{x})\|$ for any $\mathbf{x} \in U$. Since L is the limit point set of $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$, it follows that $\mathbf{x}^{(n)} \in U$ when n is large enough. So there exists a positive integer l such that $|f(\mathbf{x}^{(n)}) - \underline{f}|^{\theta} \leq \|\nabla f(\mathbf{x}^{(n)})\|$ when $n \geq l$.

The concavity of the function $g(y) = (y - \underline{f})^{1-\theta}$ for some $0 < \theta < 1$ when $y \geq \underline{f}$ implies that

$$\frac{(f(\mathbf{x}^{(n)}) - \underline{f})^{1-\theta} - (f(\mathbf{x}^{(n+1)}) - \underline{f})^{1-\theta}}{f(\mathbf{x}^{(n)}) - f(\mathbf{x}^{(n+1)})} \geq (1-\theta)(f(\mathbf{x}^{(n)}) - \underline{f})^{-\theta}.$$

Since

$$f(\mathbf{x}^{(n)}) - f(\mathbf{x}^{(n+1)}) \geq \frac{1}{2}\lambda\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\|^2 \qquad \text{and}$$
$$(f(\mathbf{x}^{(n)}) - \underline{f})^{\theta} \leq \|\nabla f(\mathbf{x}^{(n)})\| \leq d\|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\|,$$

we have that

$$\frac{2d((f(\mathbf{x}^{(n)}) - \underline{f})^{1-\theta} - (f(\mathbf{x}^{(n+1)}) - \underline{f})^{1-\theta})}{(1-\theta)\lambda} \geq \frac{\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\|^2}{\|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\|}.$$

Denote $\frac{2d((f(\mathbf{x}^{(n)}) - \underline{f})^{1-\theta} - (f(\mathbf{x}^{(m)}) - \underline{f})^{1-\theta})}{(1-\theta)\lambda}$ by $e_{n,m}$ where $m \geq n$. Then,

$$\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\|^2 \leq \|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\| e_{n,n+1},$$
$$2\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| \leq \|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\| + e_{n,n+1}.$$

Thus,

$$\begin{aligned} 2\sum_{n=l}^{k} \|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| &\leq \sum_{n=l}^{k} \|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\| + \sum_{n=l}^{k} e_{n,n+1} \\ &\leq \sum_{n=l}^{k} \|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| + \|\mathbf{x}^{(l)} - \mathbf{x}^{(l-1)}\| + \sum_{n=l}^{k} e_{n,n+1} \\ &= \sum_{n=l}^{k} \|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| + \|\mathbf{x}^{(l)} - \mathbf{x}^{(l-1)}\| + e_{l,k+1}. \end{aligned}$$

So, $\sum_{n=l}^{k} \|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| \leq \|\mathbf{x}^{(l)} - \mathbf{x}^{(l-1)}\| + e_{l,k+1}$. Since $\lim_{n\to\infty} \|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| = 0$ and $e_{l,k+1}$ is bounded for any $k \geq l$, $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$ is a Cauchy sequence. Hence, $\lim_{n\to\infty} \mathbf{x}^{(n)} = \mathbf{x}^*$ and $\nabla f(\mathbf{x}^*) = 0$. □

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The proof here can also be shown by using the techniques in [1] since the RALS algorithm satisfies the strong descent conditions of analytic cost functions. As shown in [2, 5], the global convergence rate can be further discussed regarding the value of θ. In particular, $\theta \in (0, 1/2]$ gives a linear global convergent rate while $\theta \in (1/2, 1)$ leads to a sublinear one. But there is no further information on the specific value of θ for the residual functional of the RALS algorithm. In Appendix A, we discuss the local convergence rate of RALS and show that when the sequence is close enough to the local minimum point, the RALS algorithm has a linear local convergence rate.

6. Conclusions and future outlook. We discussed convergence and acceleration of the regularized alternating least-squares (RALS) algorithm for tensor approximations. Under mild conditions, the RALS algorithm achieves global convergence and a linear local convergence rate (see Appendix A). As shown in the simulation experiments, the accelerated versions of the (R)ALS algorithm provide a higher speed of convergence compared to the original ones. Although the update map T for the acceleration keeps fixed-points invariant, it still lacks a theoretical guarantee of the effectiveness of this acceleration. Moreover, we would like to understand why a faster convergent rate can be obtained by letting the regularization parameter decrease to zero. Furthermore, we are very interested in knowing if these convergence theories have any connection in generating swamps for tensor approximations.

Acknowledgements. The authors are thankful to Hedy Attouch for some valuable suggestions on some references.

Appendix A. Local convergence rate of RALS. First we introduce some basic properties of the update operator S defined in (2.6).

THEOREM A.1. *The operator S is smooth in the space $\mathscr{X} = \mathbb{R}^{I\times r} \times \mathbb{R}^{J\times r} \times \mathbb{R}^{K\times r}$. If $\mathbf{x}^* = (\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ is a local minimum point of f, then $\mathbf{x}^*$ is a fixed-point of S.*

Proof. From the update mechanism (2.5) and the expressions (2.7) for $\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}$, and $\mathbf{C}^{(n+1)}$, it follows that the update operator S is smooth in $\mathscr{X} = \mathbb{R}^{I\times r} \times \mathbb{R}^{J\times r} \times \mathbb{R}^{K\times r}$.

If $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ is a local minimum point of f, then we have that $\mathbf{A}^{(n+1)} = \mathbf{A}^*$ when $\mathbf{B}^{(n)} = \mathbf{B}^*, \mathbf{C}^{(n)} = \mathbf{C}^*$. Since $f(\mathbf{A}, \mathbf{B}^*, \mathbf{C}^*) + \frac{1}{2}\lambda\|\mathbf{A} - \mathbf{A}^*\|^2$ is a strict convex function in $\mathbf{A}$, it follows from the update mechanism shown in (2.5) that

$$f(\mathbf{A}^{(n+1)}, \mathbf{B}^*, \mathbf{C}^*) + \frac{1}{2}\lambda\|\mathbf{A}^{(n+1)} - \mathbf{A}^*\|^2 < f(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$$

if $\mathbf{A}^{(n+1)} \neq \mathbf{A}^*$. Thus, $f(\mathbf{A}^{(n+1)}, \mathbf{B}^*, \mathbf{C}^*) < f(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$. Since f is a convex function in $\mathbf{A}$ when fixing $\mathbf{B}, \mathbf{C}$, we obtain that $f(a\mathbf{A}^{(n+1)} + (1-a)\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*) < f(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ for any $a \in (0, 1)$, which contradicts the fact that $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ is a local minimum of f. So if $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ is a local minimum point of f, we have that $\mathbf{A}^{(n+1)} = \mathbf{A}^*$ when $\mathbf{B}^{(n)} = \mathbf{B}^*$, $\mathbf{C}^{(n)} = \mathbf{C}^*$. Furthermore, it follows that $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*) = S(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ from (2.5). Thus, a local minimum point $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ of f is a fixed-point of S. ∎

Next, we discuss the contractive property of the operator S within the framework of iterative methods for nonlinear equations [23]. A similar approach [25, 26] has been applied for the ALS algorithm as well as for the alternating linear scheme for the tensor train format [12].

Any point $\mathbf{x} = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \in \mathbb{R}^{I\times r} \times \mathbb{R}^{J\times r} \times \mathbb{R}^{K\times r}$ can be viewed as a vector $\mathbf{x} = (\mathbf{x}_A^T, \mathbf{x}_B^T, \mathbf{x}_C^T)^T$, where $\mathbf{x}_A \in \mathbb{R}^{rI}$, $\mathbf{x}_B \in \mathbb{R}^{rJ}$, $\mathbf{x}_C \in \mathbb{R}^{rK}$ are the vectorized form (column stacked) of $\mathbf{A}, \mathbf{B}, \mathbf{C}$, respectively. Define the vector value function,

$$g_A(\mathbf{x}, \mathbf{y}) := \frac{\partial f(\mathbf{x}_A, \mathbf{y}_B, \mathbf{y}_C)}{\partial \mathbf{x}_A} + \lambda(\mathbf{x}_A - \mathbf{y}_A),$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

where $\mathbf{y} = (\mathbf{y}_A^T, \mathbf{y}_B^T, \mathbf{y}_C^T)^T$ and $\mathbf{y}_A \in \mathbb{R}^{rI}$, $\mathbf{y}_B \in \mathbb{R}^{rJ}$, $\mathbf{y}_C \in \mathbb{R}^{rK}$, and similarly, define

$$g_B(\mathbf{x}, \mathbf{y}) := \frac{\partial f(\mathbf{x}_A, \mathbf{x}_B, \mathbf{y}_C)}{\partial \mathbf{x}_B} + \lambda(\mathbf{x}_B - \mathbf{y}_B)$$
$$g_C(\mathbf{x}, \mathbf{y}) := \frac{\partial f(\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C)}{\partial \mathbf{x}_C} + \lambda(\mathbf{x}_C - \mathbf{y}_C).$$

Denote the vector value function $\big(g_A^T(\mathbf{x}, \mathbf{y}), g_B^T(\mathbf{x}, \mathbf{y}), g_C^T(\mathbf{x}, \mathbf{y})\big)^T$ by $G(\mathbf{x}, \mathbf{y})$. From the equations in (5.2), we know that $G(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n)}) = 0$.

Let $\mathbf{x}^*$ be a local minimizer of the residual functional f. Since f is a twice continuously differentiable function, the Hessian matrix $\mathbf{H} = \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x} \partial \mathbf{x}}$ of f at $\mathbf{x}^*$ is positive semidefinite and has nine block matrices corresponding to $\mathbf{A}, \mathbf{B}, \mathbf{C}$. From a direct computation we observe that the matrix $\frac{\partial G(\mathbf{x}^*, \mathbf{x}^*)}{\partial \mathbf{x}}$ is the lower triangular block matrix of $\mathbf{H}$ with an additional $\lambda\mathbf{I}$ on the diagonal blocks and the matrix $\frac{\partial G(\mathbf{x}^*, \mathbf{x}^*)}{\partial \mathbf{y}}$ is the strictly upper block matrix of $\mathbf{H}$ minus $\lambda\mathbf{I}$, where $\mathbf{I}$ is the identity matrix in $\mathbb{R}^{r(I+J+K)\times r(I+J+K)}$. The structure is given in the following formulas:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_A \partial \mathbf{x}_A} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_B \partial \mathbf{x}_A} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_C \partial \mathbf{x}_A} \\ \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_A \partial \mathbf{x}_B} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_B \partial \mathbf{x}_B} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_C \partial \mathbf{x}_B} \\ \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_A \partial \mathbf{x}_C} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_B \partial \mathbf{x}_C} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_C \partial \mathbf{x}_C} \end{bmatrix},$$

$$\frac{\partial G(\mathbf{x}^*, \mathbf{x}^*)}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_A \partial \mathbf{x}_A} + \lambda\mathbf{I}_A & \mathbf{0} & \mathbf{0} \\ \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_A \partial \mathbf{x}_B} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_B \partial \mathbf{x}_B} + \lambda\mathbf{I}_B & \mathbf{0} \\ \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_A \partial \mathbf{x}_C} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_B \partial \mathbf{x}_C} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_C \partial \mathbf{x}_C} + \lambda\mathbf{I}_C \end{bmatrix},$$

$$\frac{\partial G(\mathbf{x}^*, \mathbf{x}^*)}{\partial \mathbf{y}} = \begin{bmatrix} -\lambda\mathbf{I}_A & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_B \partial \mathbf{x}_A} & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_C \partial \mathbf{x}_A} \\ \mathbf{0} & -\lambda\mathbf{I}_B & \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x}_C \partial \mathbf{x}_B} \\ \mathbf{0} & \mathbf{0} & -\lambda\mathbf{I}_C \end{bmatrix},$$

where $\mathbf{I}_A, \mathbf{I}_B, \mathbf{I}_C$ are identity matrices in $\mathbb{R}^{rI\times rI}, \mathbb{R}^{rJ\times rJ}, \mathbb{R}^{rK\times rK}$, respectively.

The matrix $\frac{\partial G(\mathbf{x}^*, \mathbf{x}^*)}{\partial \mathbf{x}}$ is nonsingular since all the three diagonal blocks of $\mathbf{H}$ are positive semidefinite. The Hessian matrix $\mathbf{H}$ can be rewritten into $\mathbf{D} - \mathbf{L} - \mathbf{U}$, where $\mathbf{D}$ is a diagonal block matrix, $-\mathbf{L}$ is a strictly lower block matrix and $-\mathbf{U}$ is a strictly upper block matrix of $\mathbf{H}$. Thus we have that

$$-\frac{\partial G(\mathbf{x}^*, \mathbf{x}^*)}{\partial \mathbf{x}}^{-1} \frac{\partial G(\mathbf{x}^*, \mathbf{x}^*)}{\partial \mathbf{y}} = (\lambda\mathbf{I} + \mathbf{D} - \mathbf{L})^{-1}(\lambda\mathbf{I} + \mathbf{U})$$
$$= I - (\lambda\mathbf{I} + \mathbf{D} - \mathbf{L})^{-1}(\mathbf{D} - \mathbf{L} - \mathbf{U}).$$

Let $\mathbf{M} = \lambda\mathbf{I} + \mathbf{D} - \mathbf{L}$. From[15, Theorem 3.2], since $\mathbf{M} + \mathbf{M}^T - \mathbf{H}$ is positive definite, it follows that $\|\mathbf{I} - \mathbf{M}^{-1}\mathbf{H}\|_{\mathbf{H}} = \max\limits_{\|\mathbf{x}\|_{\mathbf{H}}\neq 0} \frac{\|(\mathbf{I}-\mathbf{M}^{-1}\mathbf{H})\mathbf{x}\|_{\mathbf{H}}}{\|\mathbf{x}\|_{\mathbf{H}}} < 1$, where $\|\mathbf{y}\|_{\mathbf{H}} = (\mathbf{y}^T\mathbf{H}\mathbf{y})^{\frac{1}{2}}$ is a seminorm on $\mathbf{y}$. If we further assume that $\mathbf{H}$ is a positive definite matrix, $\|\mathbf{y}\|_{\mathbf{H}}$ is a norm of $\mathbf{y}$ and $\|\mathbf{I} - \mathbf{M}^{-1}\mathbf{H}\|_{\mathbf{H}}$ is a matrix norm of $\mathbf{I} - \mathbf{M}^{-1}\mathbf{H}$.

Since $\mathbf{x}^*$ is a local minimum point of f, we have that $\mathbf{x}^*$ is a fixed-point of S from Theorem A.1. Furthermore, it follows that $G(\mathbf{x}^*, \mathbf{x}^*) = 0$ by the equations in (5.2). Then from the implicit function theorem, there is a neighborhood U of $\mathbf{x}^*$ such that $\mathbf{x} = S(\mathbf{y})$ when $\mathbf{y} \in U$ and $S'(\mathbf{x}^*) = \mathbf{I} - \mathbf{M}^{-1}\mathbf{H}$. Since $S'(\mathbf{x}^*) = \mathbf{I} - \mathbf{M}^{-1}\mathbf{H}$ and $\|S'(\mathbf{x}^*)\|_{\mathbf{H}} < q < 1$,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

there exists a small enough neighborhood V of $\mathbf{x}^*$ such that $\|S'(\mathbf{y})\|_{\mathbf{H}} < q$ for $\mathbf{y} \in V$. So, there exists a sufficiently small neighborhood W of $\mathbf{x}^*$ such that

$$S(\mathbf{y}) \in W, \quad \|S'(\mathbf{y})\|_{\mathbf{H}} < q, \quad \text{and} \quad \|S(\mathbf{y}) - S(\mathbf{x}^*)\|_{\mathbf{H}} < q\|\mathbf{y} - \mathbf{x}^*\|_{\mathbf{H}} \quad \text{for } \mathbf{y} \in W.$$

Hence, if $\mathbf{x}^{(n)} \in W$ for some $n \in \mathbb{N}$, then

$$\mathbf{x}^{(n+1)} \in W, \qquad \|\mathbf{x}^{(n+1)} - \mathbf{x}^*\|_{\mathbf{H}} < q\|\mathbf{x}^{(n)} - \mathbf{x}^*\|_{\mathbf{H}}, \qquad \text{and } \lim_{n\to\infty} \mathbf{x}^{(n)} = \mathbf{x}^*.$$

Furthermore, if $\mathbf{x}^{(0)} \in W$, then we obtain that

$$\limsup_{n\to\infty} \|\mathbf{x}^{(n)} - \mathbf{x}^*\|^{\frac{1}{n}} \le q, \qquad \text{and } \limsup_{n\to\infty} \|f(\mathbf{x}^{(n)}) - f(\mathbf{x}^*)\|^{\frac{1}{n}} \le q$$

from the equivalence of norms in the finite-dimensional space. So we obtain that the RALS algorithm has a linear local convergence rate when $\mathbf{x}^n$ is enough close to a local minimum point $\mathbf{x}^*$ and the Hessian matrix $\frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x} \partial \mathbf{x}}$ of f at $\mathbf{x}^*$ is positive definite.

THEOREM A.2. *Let $\{\mathbf{x}^{(n)}\}_{n\in\mathbb{N}}$ be the sequence generated by the RALS algorithm. Assume that $\mathbf{x}^*$ is a local minimum point of f and the Hessian matrix $H = \frac{\partial^2 f(\mathbf{x}^*)}{\partial \mathbf{x} \partial \mathbf{x}}$ is positive definite. There exist a neighborhood W of $\mathbf{x}^*$ and a positive constant $q < 1$ such that:*

1. *If $\mathbf{x}^{(n)} \in W$ for some $n \in \mathbb{N}$, then $\mathbf{x}^{(n+1)} \in W$, $\|\mathbf{x}^{(n+1)} - \mathbf{x}^*\|_{\mathbf{H}} < q\|\mathbf{x}^{(n)} - \mathbf{x}^*\|_{\mathbf{H}}$ and $\lim_{n\to\infty} \mathbf{x}^{(n)} = \mathbf{x}^*$.*
2. *If $\mathbf{x}^{(0)} \in W$, then $\limsup_{n\to\infty} \|\mathbf{x}^{(n)} - \mathbf{x}^*\|^{\frac{1}{n}} \le q$ and $\limsup_{n\to\infty} \|f(\mathbf{x}^{(n)}) - f(\mathbf{x}^*)\|^{\frac{1}{n}} \le q$.*

In the work of Uschmajew [26], a similar result was provided for the ALS algorithm with the objective functional $g_\lambda(\mathbf{A}, \mathbf{B}, \mathbf{C}) = f(\mathbf{A}, \mathbf{B}, \mathbf{C}) + \lambda(\|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 + \|\mathbf{C}\|^2)$. Naturally, a large enough λ yields positive definiteness of $\frac{\partial^2 g_\lambda(\mathbf{x}^*)}{\partial \mathbf{x} \partial \mathbf{x}}$ which guarantees a linear convergent rate.

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), 531–547.

[2] H. ATTOUCH, J. BOLTE, P. REDONT, AND A. SOUBEYRAN, *Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457.

[3] H. ATTOUCH, J. BOLTE, AND B. F. SVAITER, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program., 137 (2013), pp. 91–129.

[4] A. BECK AND M. TEBOULLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[5] J. BOLTE, S. SABACH, AND M. TEBOULLE, *Proximal alternating linearized minimization nonconvex and nonsmooth problems*, Math. Program., 146 (2014), pp. 459–494.

[6] 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.

[7] B. CHEN, S. HE, Z. LI, AND S. ZHANG, *Maximum block improvement and polynomial optimization*, SIAM J. Optim., 22 (2012), pp. 87–107.

[8] 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.

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

[10] A. HARAUX, *Some applications of the Łojasiewicz gradient inequality*, Commun. Pure Appl. Anal., 11 (2012), pp. 2417–2427.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[11] R. A. HARSHMAN, *Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-code factor analysis,* UCLA Working Papers in Phonetics 16, UCLA Linguistics Department, UCLA, Los Angeles, 1970 (84 pages).

[12] S. HOLTZ, T. ROHWEDDER AND R. SCHNEIDER, *The alternating linear scheme for tensor optimization in the tensor train format*, SIAM J. Sci. Comput., 34 (2012), pp. A683–A713.

[13] E. ISAACSON AND H. B. KELLER, *Analysis of Numerical Methods*, Wiley, New York, 1966.

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

[15] Y.-J. LEE, J. WU, J. XU, AND L. ZIKATANOV, *On the convergence of iterative methods for semidefinite linear systems*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 634–641.

[16] N. LI, S. KINDERMANN, AND C. NAVASCA, *Some convergent results of the regularized alternating least-squares for tensor decomposition*, Linear Algebra Appl., 438 (2013), pp. 796–812.

[17] Z. LI, A. USCHMAJEW, AND S. ZHANG, *On convergence of the maximum block improvement method*, SIAM J. Optim., 25 (2015), pp. 210–233.

[18] L.-H. LIM AND P. COMON, *Nonnegative approximations of nonnegative tensors*, J. Chemometrics, 23 (2009), pp. 432–441.

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

[20] C. NAVASCA, L. DE LATHAUWER, AND S. KINDERMANN, *Swamp reducing technique for tensor decomposition*, in Proceedings of the European Signal Processing Conference 2008, EUSIPCO online proceedings, EURASIP, 2008 (5 pages).

[21] Y. NESTEROV, *Introductory Lectures on Convex Optimization*, Springer, New York, 2004.

[22] T. NODA, *The Steffensen iteration method for systems of nonlinear equations*, Proc. Japan Acad. Ser. A Math. Sci., 60 (1984), pp. 18–21.

[23] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[24] P. PAATERO, *A weighted non-negative least squares algorithm for three-way PARAFAC factor analysis*, Chemometrics Intell. Lab. Syst., 38 (1997), pp. 223–242.

[25] T. ROHWEDDER AND A. USCHMAJEW, *On local convergence of alternating schemes for optimization of convex problems in the tensor train format*, SIAM J. Numer. Anal., 51 (2013), pp. 1134–1162.

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

[27] Y. XU AND W. YIN, *A block coordinate descent method for regularized multi-convex optimization with applications to nonnegative tensor factorization and completion*, SIAM J. Imaging Sci., 6 (2013), pp. 1758–1789.