SIAM J. MATRIX ANAL. APPL.
Vol. 33, No. 2, pp. 639–652

LOCAL CONVERGENCE OF THE ALTERNATING LEAST SQUARES ALGORITHM FOR CANONICAL TENSOR APPROXIMATION*

ANDRÉ USCHMAJEW†

Abstract. A local convergence theorem for calculating canonical low-rank tensor approximations (PARAFAC, CANDECOMP) by the alternating least squares algorithm is established. The main assumption is that the Hessian matrix of the problem is positive definite modulo the scaling indeterminacy. A discussion, whether this is realistic, and numerical illustrations are included. Also regularization is addressed.

Key words. ALS, low-rank approximation, nonlinear Gauss–Seidel, PARAFAC

AMS subject classifications. 15A69, 65K10, 65F30

DOI. 10.1137/110843587

1. Introduction.

According to the review article of Kolda and Bader [11], the alternating least squares (ALS) algorithm is still the "workhorse" in computing low-rank approximations and decompositions of high-order tensors. It has been widely used in such fields as psychometrics, chemometrics, and signal processing (see the references in [11]). The reason for the popularity of this algorithm lies in the fact that it is very simple, conceptually and numerically, while still delivering astonishingly good results in many cases, if employed with care [25].

In the present paper we investigate the ALS algorithm for low-rank approximation by tensors in the canonical format (also known as CP, PARAFAC, or CANDECOMP). The great majority of the literature focuses on global properties of the iteration, like the existence of convergent subsequences and critical points, or the occurrence of swamps [3, 6, 7, 15, 17, 21, 22]. But any widely used algorithm should desirably also be backed by a local convergence theory. To prevent any misunderstandings, by a local convergence theory we mean a theory for the parameters (iterates) of an algorithm, not for the residuals (loss function).

Surprisingly, there are few works in this direction, an exception being the work of Zhang and Golub [27] on the rank-one approximation. For higher ranks most of the difficulties with the global behavior of ALS seem to be intimately related to the fact that the approximation problem itself can be ill-posed or ill-conditioned [4]. We will not enter into this discussion in the present paper, but rather assume that a local minimum exists, since otherwise the question of local convergence does not make much sense.

The ALS algorithm is an alternating optimization scheme and as such a so-called nonlinear block Gauss–Seidel method. There is a well-developed local theory for this type of method [1, 18, 23]. It can be shown that it, up to higher-order terms, locally equals the linear block Gauss–Seidel iteration applied to the Hessian matrix at the solution. Thus it is locally linearly convergent (at the same asymptotic rate

*Received by the editors August 8, 2011; accepted for publication (in revised form) by T. G. Kolda April 11, 2012; published electronically June 28, 2012. This research was supported in part by DFG (Deutsche Forschungsgemeinschaft) through BMS (Berlin Mathematical School) and MATHEON.
http://www.siam.org/journals/simax/33-2/84358.html
†Institut für Mathematik, Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (uschmajew@math.tu-berlin.de).

as the linear Gauss–Seidel), provided that this Hessian matrix is positive definite. The problem is that local minima in canonical low-rank tensor approximations do not have this property due to the nonuniqueness of the representation caused by the scaling indeterminacy. (The permutation and sign indeterminacies are irrelevant in a local theory.) On the other hand, it is known that the Gauss–Seidel method for only semidefinite linear systems still is convergent, but only up to elements in the null space of the system matrix [9, 14]. One, hence, has to remove the null space of the Hessian from the iteration.

In this paper we present a local convergence result (Theorem 3.3) for ALS (Algorithm 1) under the assumption that the Hessian of the loss function at the solution is essentially positive definite, except on a trivial null space caused by the scaling indeterminacy (Assumption 1 in section 3.1). This assumption necessarily requires the local essential uniqueness of the CP decomposition (that is, uniqueness up to scaling), which is reasonable for tensors of order higher than three [3, 8, 12, 24]. Second-order tensors (matrices) can meet this condition at best, if they have rank one. The main idea of the proof is to show that an inbuilt normalization procedure in the ALS algorithm acts (in first order) as a projection onto a subspace complementary to the null space of the Hessian. On this subspace the linear Gauss–Seidel method then is known to be contractive. The main assumption is discussed in detail in section 3.4, and some effort is made to prove a sufficient condition for its validity in the important case of rank-one approximation (Theorem 3.6). In section 3.5, we discuss the advantages of regularization. The main ideas are not specifically related to the least squares error as the loss function to be minimized. We will consider arbitrary loss functions in section 3.3. It turns out that the global minimization in one ALS direction can be replaced by a single Newton step to obtain an approximate scheme which is still convergent (under the same assumptions).

One convenient aspect of our approach is that it avoids the explicit use of Lagrange multipliers. It is therefore easily accessible and in principle applicable to more delicate types of redundancies as they, for instance, occur in the Tucker format or in the newly developed TT format [19, 20]. This will be elaborated elsewhere.

Generalization to the complex case is not completely straightforward. The problem is that it is subtle to define a format which removes the scaling indeterminacy. We will comment on this issue at the related points of our exposition.

Notation. We use the notation $f'(\mathbf{x})$ for the derivative of a function f at $\mathbf{x}$ and $f''(\mathbf{x})$ for the Hessian. By $(\cdot,\cdot)$ and $\|\cdot\|$ we denote the Euclidian (Frobenius) inner product and norm, respectively. However, we will write $f''(\mathbf{x})[\mathbf{h},\mathbf{h}]$ instead of $(\mathbf{h}, f''(\mathbf{x})\mathbf{h})$ for the application of the second derivtive, since the former notation is also meaningful for vector-valued functions such as τ introduced later on. Note that for rank-one tensors it holds that

$$(a \otimes b \otimes c, a' \otimes b' \otimes c') = (a,a')(b,b')(c,c'). \tag{1.1}$$

2. The ALS algorithm. For the sake of clarity, we restrict ourselves most of the time to third-order tensors. The reasoning for the higher-order case is completely analogous.

Let $n_1, n_2, n_3 \in \mathbb{N} \setminus \{1\}$, and let $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $\mathcal{T} \neq 0$, be a real third-order tensor, treated here as a three-dimensional array. Given $r \in \mathbb{N}$, let

$$\mathcal{X} = \mathbb{R}^{n_1 \times r} \times \mathbb{R}^{n_2 \times r} \times \mathbb{R}^{n_3 \times r}.$$

The elements of $\mathcal{X}$ will be denoted by $\mathbf{x} = (\mathbf{A}, \mathbf{B}, \mathbf{C})$. Consider the function

$$f\colon \mathcal{X} \to \mathbb{R}\colon \mathbf{x} = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \mapsto \frac{1}{2}\left\| \mathcal{T} - \sum_{j=1}^{r} a_j \otimes b_j \otimes c_j \right\|^2, \tag{2.1}$$

where a_j, b_j, and c_j are supposed to be the columns of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, respectively. The matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are called factor matrices in the literature. We seek a solution of

$$f(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \min. \tag{2.2}$$

It is assumed that at least one local minimum of (2.2) exists. It will be denoted by $\mathbf{x}^* = (\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$.

The ALS algorithm is a simple method for (hopefully) solving (2.2). Given a starting point $\mathbf{x}^{(0)}$, it consists of iterating the cycle

$$\begin{aligned} \mathbf{A}^{(n+1)} &= \operatorname*{argmin}_{\mathbf{A} \in \mathbb{R}^{n_1 \times r}} f(\mathbf{A}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)}), \\ \mathbf{B}^{(n+1)} &= \operatorname*{argmin}_{\mathbf{B} \in \mathbb{R}^{n_2 \times r}} f(\mathbf{A}^{(n+1)}, \mathbf{B}, \mathbf{C}^{(n)}), \\ \mathbf{C}^{(n+1)} &= \operatorname*{argmin}_{\mathbf{C} \in \mathbb{R}^{n_3 \times r}} f(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}). \end{aligned} \tag{2.3}$$

This algorithm is a particular example of the nonlinear block Gauss–Seidel (relaxation) method [18, 23]. The name ALS stems from the fact that each microiteration step in (2.3) is a linear least squares problem. An explicit solution formula for the microsteps (invoking a pseudoinverse) can be given in terms of the Khatri–Rao product [11, section 3.4]. If every such step possesses a unique solution,[1] then one loop (2.3) defines an operator S via

$$(\mathbf{A}^{(n+1)}, \mathbf{B}^{(n+1)}, \mathbf{C}^{(n+1)}) = \mathbf{x}^{(n+1)} = S(\mathbf{x}^{(n)}) = S(\mathbf{A}^{(n)}, \mathbf{B}^{(n)}, \mathbf{C}^{(n)}). \tag{2.4}$$

From now on we will consider only local minima of (2.2) in the open subset

$$\hat{\mathcal{X}} = \{(\mathbf{A}, \mathbf{B}, \mathbf{C}) \in \mathcal{X} \mid a_j \neq 0,\ b_j \neq 0,\ c_j \neq 0 \text{ for } j = 1, 2, \ldots, r\}.$$

We assume such minima to exist. Restricting to $\hat{\mathcal{X}}$ is reasonable to avoid pseudoinverses, since only if $\mathbf{x}^* \in \hat{\mathcal{X}}$ we can hope (2.3) to have unique solutions (take, for instance, $c_1^{(n)} = 0$ in the first line of (2.3)). We consider local minima in $\mathcal{X} \setminus \hat{\mathcal{X}}$ to be too degenerate for our framework. For them, at least one rank-one term vanishes, so the rank parameter r should be adjusted. However, it seems to be a difficult question whether such local minima can really exist if, say, the canonical rank of the target tensor $\mathcal{T}$ is larger than or equal to r.

The major difficulty in the analysis of algorithm (2.3) lies in the fact that $\mathbf{x}^*$ cannot be an isolated local minimum of (2.2), since every rank-one term $a_j^* \otimes b_j^* \otimes c_j^*$ may be replaced by $(\alpha_j a_j^*) \otimes (\beta_j b_j^*) \otimes (\gamma_j c_j^*)$ as long as $\alpha_j \beta_j \gamma_j = 1$. If α, β, and γ are positive, we will call this operation a *rescaling* of $\mathbf{x}^*$. In fact, every such rescaled solution itself is a local minimum of (2.2) and a fixed point of the iteration (2.4).

[1] If not, one could apply a pseudoinverse or choose a solution by any other rule. The ALS iteration might be sensitive to this choice, but that is not the subject of the present paper.

So there is no reason why the iteration should, if at all, converge to a particular prescribed solution $\mathbf{x}^*$.

Additionally, when applying the ALS algorithm in the naive form (2.3), it can happen that a component, say $a_1^{(n)}$, tends to infinity while another, say $b_1^{(n)}$, compensates this by tending to zero, such that $a_1^{(n)} \otimes b_1^{(n)} \otimes c_1^{(n)}$ remains bounded. This deteriorates the condition of each microstep.

For both reasons a normalization strategy has to be invoked. The usual way to do this is to represent tensors in the form

$$\sum_{j=1}^{r} \sigma_j a_j \otimes b_j \otimes c_j \quad \text{with} \quad \|a_j\| = \|b_j\| = \|c_j\| = 1,\ \sigma_j \in \mathbb{R} \quad \text{for } j = 1, 2, \dots, r. \tag{2.5}$$

To avoid the additional parameters σ_j, we will instead consider tensors in the equilibrated format:

$$\sum_{j=1}^{r} a_j \otimes b_j \otimes c_j \quad \text{with} \quad \|a_j\| = \|b_j\| = \|c_j\| \quad \text{for } j = 1, 2, \dots, r. \tag{2.6}$$

This fixes the representation among all possible rescalings up to a change of signs.[2]

The rescaling of a tensor into the equilibrated format (2.6), *without* changing the signs of the vectors, defines an operator $R(\mathbf{x}) = R(\mathbf{A}, \mathbf{B}, \mathbf{C})$ for the corresponding parametrization $\mathbf{x}$ via

$$(a_j, b_j, c_j) \mapsto \left(\frac{\delta_j a_j}{\|a_j\|}, \frac{\delta_j b_j}{\|b_j\|}, \frac{\delta_j c_j}{\|c_j\|} \right), \quad \delta_j = (\|a_j\|\|b_j\|\|c_j\|)^{1/3}, \quad j = 1, 2, \dots, r.$$

Note that R can be defined on the whole space $\mathcal{X}$ by continuous extension (on $\mathcal{X} \setminus \hat{\mathcal{X}}$ it is zero), but is smooth only on $\hat{\mathcal{X}}$. We will call a representation $\mathbf{x} = (\mathbf{A}, \mathbf{B}, \mathbf{C})$ *equilibrated* if $R(\mathbf{x}) = \mathbf{x}$. This can happen only in $\hat{\mathcal{X}}$ or at the origin.

The ALS algorithm for calculating an equilibrated local solution of (2.2) reads as follows.

ALGORITHM 1. ALS with equilibration.

Require: $\mathbf{x}^{(0)} = (\mathbf{A}^{(0)}, \mathbf{B}^{(0)}, \mathbf{C}^{(0)})$
for $n = 0, 1, 2, \dots$ **do**
1. Perform one ALS cycle:
   $\tilde{\mathbf{x}}^{(n+1)} = S(\mathbf{x}^{(n)})$.
2. Equilibrate the factor matrices:
   $\mathbf{x}^{(n+1)} = R(\tilde{\mathbf{x}}^{(n+1)})$.

end for

Fixed points of this algorithm are necessarily equilibrated. Other variants of ALS are possible which, for instance, include normalization of the iterates in the sense of (2.5) instead of equilibration (as presented in [11]). To increase the numerical stability it may also be reasonable to equilibrate after each microstep of (2.3). All such variants of the algorithm are equivalent in the sense that they, given the same starting point, produce the same sequence of iterates up to rescaling. The presented

[2] In the complex case, up to rotations on the unit sphere.

version with equilibration is favorable for the analysis, but the convergence result, as stated in Theorem 3.3, trivially transfers to different scaling strategies. In fact, in our numerical experiments we used normalized iterates of the form (2.5).

3. Convergence analysis. From now on $\mathbf{x}^* = (\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ will always denote a nonzero equilibrated local solution of (2.2). Recall then that $\mathbf{x}^* \in \hat{\mathcal{X}}$. In this section we establish a local convergence theorem for Algorithm 1 in a neighborhood of $\mathbf{x}^*$.

3.1. The positive definiteness assumption. We first return our attention to the scaling indeterminacy again. The function f in (2.1) is constant on the $2r$-dimensional (in the real case not connected) submanifold[3]

$$\mathcal{M}^* = \{(\mathbf{A}^*\Delta_1, \mathbf{B}^*\Delta_2, \mathbf{C}^*\Delta_1^{-1}\Delta_2^{-1}) \in \mathcal{X} \mid \Delta_1, \Delta_2 \text{ nonsingular diagonal matrices}\},$$

which contains all representations of $\sum_{j=1}^r a_j^* \otimes b_j^* \otimes c_j^*$ that can be obtained by rescaling (including sign changing). Since every point in $\mathcal{M}^*$ is a local minimum of (2.2), the derivative f' vanishes on $\mathcal{M}^*$. Consequently, the Hessian $f''(\mathbf{x}^*)$ has at most rank $\dim \mathcal{X} - 2r = r(n_1 + n_2 + n_3) - 2r$. More precisely, let $T\mathcal{M}^*_{\mathbf{x}^*}$ denote the tangent space of $\mathcal{M}^*$ at $\mathbf{x}^*$, then $f''(\mathbf{x}^*)[\mathbf{h}, \mathbf{h}] = 0$ for all $\mathbf{h} \in T\mathcal{M}^*_{\mathbf{x}^*}$. It is, by the way, easy to see that

$$T\mathcal{M}^*_{\mathbf{x}^*} = \{(\mathbf{A}^*\Delta_1, \mathbf{B}^*\Delta_2, -\mathbf{C}^*(\Delta_1 + \Delta_2)) \in \mathcal{X} \mid \Delta_1, \Delta_2 \text{ diagonal matrices}\}. \tag{3.1}$$

We now make the following assumption.

ASSUMPTION 1. *The rank of $f''(\mathbf{x}^*)$ equals $r(n_1 + n_2 + n_3) - 2r$, that is, the null space of $f''(\mathbf{x}^*)$ is $T\mathcal{M}^*_{\mathbf{x}^*}$.*

In other words, $f''(\mathbf{x}^*)$ shall be positive definite in every direction except those tangent to the rescalings. This implies that the parametrization $\mathbf{x}^*$ is locally essentially unique (the converse may not always be true). We will discuss in section 3.4 whether Assumption 1 is realistic. In any case, it seems unavoidable for a standard local convergence proof of Algorithm 1 for the following reason.

LEMMA 3.1. *$R'(\mathbf{x}^*)$ is a projector whose null space is precisely $T\mathcal{M}^*_{\mathbf{x}^*}$.*

Proof. First of all, $R = R \circ R$ shows that

$$R'(\mathbf{x}^*) = R'(R(\mathbf{x}^*))R'(\mathbf{x}^*) = R'(\mathbf{x}^*)R'(\mathbf{x}^*)$$

is a projector. Since R is constant on the connected components[4] of $\mathcal{M}^*$, it also follows that $R'(\mathbf{x}^*)\mathbf{h} = 0$ for all $\mathbf{h} \in T\mathcal{M}^*_{\mathbf{x}^*}$.

On the other hand, R is the identity on the set of all equilibrated $\mathbf{x}$, in particular on the submanifold of all $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ whose columns have the same norm as the corresponding columns of $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$. Hence $R'(\mathbf{x}^*)\mathbf{h} = \mathbf{h}$ for all $\mathbf{h}$ from the tangent space of that submanifold at $\mathbf{x}^*$. Since the latter can be regarded as a Cartesian product of spheres (on which the columns of the matrices are located), it is clear that its tangent space at $\mathbf{x}^*$ is

$$U^* = \{(\mathbf{A}, \mathbf{B}, \mathbf{C}) \in \mathcal{X} \mid a_j \perp a_j^*,\ b_j \perp b_j^*,\ c_j \perp c_j^* \text{ for } j = 1, 2, \ldots, r\},$$

[3]The main reason why this is a submanifold of the specified dimension is that the matrices $\mathbf{A}^*$, $\mathbf{B}^*$, and $\mathbf{C}^*$ contain no zero columns. In the order-d case the corresponding submanifold is of dimension $(d-1)r$.

[4]In the complex case, $\mathcal{M}^*$ consists of only one component. One suggestion for obtaining a constant equilibration operator R is to fix certain positions in the vectors a_j and b_j, and rotate the corresponding entries to the positive real axis, assuming that they are not zero in a neighborhood of $\mathbf{x}^*$. Although such positions surely exist, this is a little bit unsatisfactory.

where $\perp$ means Euclidian orthogonality.

Finally, it is easy to see that, for instance, $R'(\mathbf{x}^*)\mathbf{h} \neq 0$ for all $\mathbf{h} \neq 0$ from

$$V^* = \{(\mathbf{A}^*\Delta, 0, 0) \in \mathcal{X} \mid \Delta \text{ diagonal matrix}\}.$$

This finishes the proof for $\dim T\mathcal{M}^*_{\mathbf{x}^*} + \dim(U^* \oplus V^*) = \dim \mathcal{X}$. ∎

Interestingly, Assumption 1 already ensures that Algorithm 1 is well-defined in a neighborhood of $\mathbf{x}^*$.

LEMMA 3.2. *Assume that Assumption* 1 *holds. Then the ALS operator* S *in* (2.4) *is well-defined and continuously differentiable in some neighborhood of* $\mathbf{x}^*$. *Moreover,* $\mathbf{x}^*$ *is a fixed point of* S *and we have* $S'(\mathbf{x}^*) = I - M^{-1}f''(\mathbf{x}^*)$, *where* M *is the lower block triangular (including the block diagonal) part of* $f''(\mathbf{x}^*)$ *corresponding to the partition* $\mathbf{x} = (\mathbf{A}, \mathbf{B}, \mathbf{C})$.

In a possibly smaller neighborhood the composition $R \circ S$ *is well-defined and continuously differentiable. Consequently, one loop of Algorithm* 1 *is feasible if the current iterate* $\mathbf{x}^{(n)}$ *is close enough to* $\mathbf{x}^*$.

Proof. The main argument is that the diagonal blocks of $f''(\mathbf{x}^*)$ are positive definite. To see this for the first block, consider $\mathbf{h} = (\mathbf{H_A}, 0, 0)$ with $\mathbf{H_A} \in \mathbb{R}^{n_1 \times r}$, $\mathbf{H_A} \neq 0$. Since, by (3.1), $\mathbf{h} \notin T\mathcal{M}^*_{\mathbf{x}^*}$, Assumption 1 guarantees $f''(\mathbf{x}^*)[\mathbf{h}, \mathbf{h}] > 0$, which shows that the first diagonal block of $f''(\mathbf{x}^*)$, corresponding to the block variable $\mathbf{A}$, is positive definite. The same reasoning works for all blocks.

It follows that the diagonal blocks of $f''(\mathbf{x})$ are positive definite for all $\mathbf{x}$ sufficiently close to $\mathbf{x}^*$. Every microstep in (2.3) is a quadratic minimization problem whose system matrix is the corresponding diagonal block of the Hessian $f''(\mathbf{x})$ at the current point. Hence, if $\mathbf{x}^{(n)}$ is close enough to $\mathbf{x}^*$, then every microstep, taken by itself, possesses a unique (global) solution, which depends smoothly on the input. Clearly, $\mathbf{x}^*$ is a fixed point of (2.3). Therefore, we can even choose a neighborhood so small that all three microsteps can be executed consecutively with unique solutions, that is, S is well-defined and smooth in this neighborhood. Since the fixed point $\mathbf{x}^*$ of S is in the open set $\hat{\mathcal{X}}$, we also have $S(\mathbf{x}^{(n)}) \in \hat{\mathcal{X}}$ if $\mathbf{x}^{(n)}$ is close enough to $\mathbf{x}^*$, that is, $(R \circ S)(\mathbf{x}^{(n)})$ is well-defined and smooth then.

For a proof that $S'(\mathbf{x}^*) = I - M^{-1}f''(\mathbf{x}^*)$, see [1, Lemma 2] or [18, 10.3.5]. ∎

As we have seen in the proof, Lemma 3.2, in principle, holds under the weaker assumption that the diagonal blocks of $f''(\mathbf{x}^*)$ are positive definite. This is equivalent to the unique solvability of each microstep (2.3) with input $\mathbf{x}^*$. For completeness we remark that this, in turn, is equivalent to the linear independence of the sets of complementary tensors

$$\{b_j^* \otimes c_j^* \mid j = 1, 2, \dots, r\}, \quad \{a_j^* \otimes c_j^* \mid j = 1, 2, \dots, r\}, \quad \{a_j^* \otimes b_j^* \mid j = 1, 2, \dots, r\}.$$

This is, for instance, the case if the solution tensor $\sum_{j=1}^r a_j^* \otimes b_j^* \otimes c_j^*$ has canonical rank r [5].

3.2. Convergence theorem. The proof of convergence goes as follows. By the contraction principle, Algorithm 1 will be linearly convergent in a neighborhood of its fixed point $\mathbf{x}^* \in \hat{\mathcal{X}}$ if the spectral radius of $(R \circ S)'(\mathbf{x}^*) = R'(\mathbf{x}^*)S'(\mathbf{x}^*)$ is less than one, that is, if $(R'(\mathbf{x}^*)S'(\mathbf{x}^*))^n \to 0$ for $n \to \infty$. Since $S'(\mathbf{x}^*)$ is the identity on the null space $T\mathcal{M}^*_{\mathbf{x}^*}$ of $R'(\mathbf{x}^*)$, this is equivalent to

$$R'(\mathbf{x}^*)(S'(\mathbf{x}^*))^n \to 0 \quad \text{for } n \to \infty. \tag{3.2}$$

By Lemma 3.2, $S'(\mathbf{x}^*)$ is the error iteration matrix of the linear block Gauss–Seidel method for $f''(\mathbf{x}^*)$. For any $\mathbf{h} \in \mathcal{X}$, the sequence $(S'(\mathbf{x}^*))^n\mathbf{h}$ is known to converge to

an element in the null space of $f''(\mathbf{x}^*)$, provided that $f''(\mathbf{x}^*)$ is positive semidefinite with a positive block diagonal [9, Theorem 2]. If Assumption 1 holds, the latter is the case (see the proof of Lemma 3.2), and, moreover, Lemma 3.1 implies (3.2).

In the following theorem we say a little more by specifying a vector norm in which $(R \circ S)'(\mathbf{x}^*)$ is a contraction. Let

$$|\mathbf{x}|_E = (f''(\mathbf{x}^*)[\mathbf{x}, \mathbf{x}])^{1/2}$$

denote the *energy seminorm* of $f''(\mathbf{x}^*)$. If Assumption 1 holds, then, by Lemma 3.1,

$$|\mathbf{x}|_*^2 = \|(I - R'(\mathbf{x}^*))\mathbf{x}\|^2 + |\mathbf{x}|_E^2$$

defines a norm on $\mathcal{X}$. One can regard this norm as the energy norm of $f''(\mathbf{x}^*)$ with the zero eigenvalues replaced by one.

THEOREM 3.3. *Let $\mathbf{x}^* = (\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ be an equilibrated local minimum of (2.2) for which Assumption 1 holds. Then for every $\epsilon > 0$ there exists a neighborhood of $\mathbf{x}^*$, such that for any starting point $\mathbf{x}^{(0)}$ in this neighborhood the iterates of Algorithm 1 converge linearly to $\mathbf{x}^*$, and particularly satisfy*

$$|\mathbf{x}^{(n+1)} - \mathbf{x}^*|_* \le (q + \epsilon)|\mathbf{x}^{(n)} - \mathbf{x}^*|_*$$

(Q-linear convergence), where $q = |S'(\mathbf{x}^*)|_E < 1$. The sequence of tensors $(\tau(\mathbf{x}^{(n)}))$ converges at least R-linearly to $\tau(\mathbf{x}^*)$ at the same asymptotic rate, that is,

$$\limsup_{n\to\infty} \|\tau(\mathbf{x}^{(n)}) - \tau(\mathbf{x}^*)\|^{1/n} \le q.$$

Note that we did not assert local convergence if $\tau(\mathbf{x}^{(0)})$ is close enough to $\tau(\mathbf{x}^*)$. Even if equilibrated, it is usually not possible to bound $\|\mathbf{x}^{(n)} - \mathbf{x}^*\|$ in terms of $\|\tau(\mathbf{x}^{(n)}) - \tau(\mathbf{x}^*)\|$.

Proof. By Lemma 3.2, the operator $R \circ S$ is well-defined and continuously differentiable in a neighborhood of its fixed point $\mathbf{x}^*$. We have to show $|(R \circ S)'(\mathbf{x}^*)|_* \le q < 1$.

It holds that $f(R(\mathbf{x}^* + \mathbf{h})) = f(\mathbf{x}^* + \mathbf{h})$ for all sufficiently small $\mathbf{h}$. Since $f'(\mathbf{x}^*) = f'(R(\mathbf{x}^*)) = 0$, it follows that

$$|R'(\mathbf{x}^*)\mathbf{h}|_E^2 = f''(\mathbf{x}^*)[R'(\mathbf{x}^*)\mathbf{h}, R'(\mathbf{x}^*)\mathbf{h}] = f''(\mathbf{x}^*)[\mathbf{h}, \mathbf{h}] = |\mathbf{h}|_E^2.$$

Additionally, by Lemma 3.1, $(I - R'(\mathbf{x}^*))R'(\mathbf{x}^*) = 0$.[5] Hence, for all $\mathbf{h} \in \mathcal{X}$ we have

$$|(R \circ S)'(\mathbf{x}^*)\mathbf{h}|_* = |R'(\mathbf{x}^*)S'(\mathbf{x}^*)\mathbf{h}|_* = |S'(\mathbf{x}^*)\mathbf{h}|_E \le |S'(\mathbf{x}^*)|_E|\mathbf{h}|_E \le |S'(\mathbf{x}^*)|_E|\mathbf{h}|_*.$$

It is known that the error iteration matrix of the linear block Gauss–Seidel method is a contraction in the energy seminorm, that is, $|S'(\mathbf{x}^*)|_E < 1$ if $f''(\mathbf{x}^*)$ is semidefinite with positive definite block diagonal; see [9, eq. (9)] or [14, Theorem 3.2]. Since the latter is ensured by Assumption 1, the first part of the theorem is proved.

As τ is Lipschitz continuous on any compact subset of $\mathcal{X}$, it follows immediately that $\limsup_{n\to\infty} \|\tau(\mathbf{x}^{(n)}) - \tau(\mathbf{x}^*)\|^{1/n} \le \limsup_{n\to\infty} |\mathbf{x}^{(n)} - \mathbf{x}^*|_*^{1/n} \le q$. ∎

For (quite involved) generic estimates of $q = |S'(\mathbf{x}^*)|_E$, we refer to [26].

3.3. General target functions. The concrete form of the function f only entered in Lemma 3.2, where we used that every microstep of the ALS iteration (2.3) is a quadratic minimization problem. Due to the multilinearity of the tensor product,

[5] Actually, $R'(\mathbf{x}^*)$ is an orthogonal projector with respect to the inner product of the norm $|\cdot|_*$.

this is the case for any quadratic cost function $J\colon \mathbb{R}^{n_1\times n_2\times n_3} \to \mathbb{R}$, so our theorem applies without change to

$$f(\mathbf{A},\mathbf{B},\mathbf{C}) = J\left(\sum_{j=1}^{r} a_j \otimes b_j \otimes c_j\right). \tag{3.3}$$

Important choices for J include energy norms of selfadjoint partial differential operators. There have been attempts to use a tensor calculus in the solution of such problems in very high dimensions, for instance, in [2].

In the case of a general nonlinear functional J, the microsteps in (2.3) do not need to have unique solutions, even if assumptions on the Hessian are made. Moreover, the global minima of a microstep might lie far away from the considered local minimum $\mathbf{x}^*$. To stay local, one could replace in each microstep the function f in (3.3) by its second-order expansion at the current microiterate and minimize that one, which means nothing other than performing a single Newton step with respect to the current block variable. This procedure is called approximate nonlinear relaxation in [23] and the nonlinear Gauss–Seidel–Newton method in [18]. Explicitly, it reads as

$$\begin{aligned}
\mathbf{A}^{(n+1)} &= \mathbf{A}^{(n)} - [f''_{\mathbf{A}}(\mathbf{A}^{(n)},\mathbf{B}^{(n)},\mathbf{C}^{(n)})]^{-1}\cdot \nabla_{\mathbf{A}} f(\mathbf{A}^{(n)},\mathbf{B}^{(n)},\mathbf{C}^{(n)}),\\
\mathbf{B}^{(n+1)} &= \mathbf{B}^{(n)} - [f''_{\mathbf{B}}(\mathbf{A}^{(n+1)},\mathbf{B}^{(n)},\mathbf{C}^{(n)})]^{-1}\cdot \nabla_{\mathbf{B}} f(\mathbf{A}^{(n+1)},\mathbf{B}^{(n)},\mathbf{C}^{(n)}),\\
\mathbf{C}^{(n+1)} &= \mathbf{C}^{(n)} - [f''_{\mathbf{C}}(\mathbf{A}^{(n+1)},\mathbf{B}^{(n+1)},\mathbf{C}^{(n)})]^{-1}\cdot \nabla_{\mathbf{C}} f(\mathbf{A}^{(n+1)},\mathbf{B}^{(n+1)},\mathbf{C}^{(n)}),
\end{aligned} \tag{3.4}$$

where we denoted by $\nabla_{\mathbf{A}} f$ and $f''_{\mathbf{A}}$ the gradient and Hessian (diagonal block) with regard to variable $\mathbf{A}$ only (same for $\mathbf{B}$ and $\mathbf{C}$). If we denote by $\hat{S}$ the iteration operator of (3.4), the claim of Lemma 3.2 and its proof hold for $\hat{S}$ (cf. [18, 10.3.3]). In fact, if J is quadratic as above, then $\hat{S} = S$.

THEOREM 3.4. *Let $J \in C^2(\mathbb{R}^{n_1\times n_2\times n_3},\mathbb{R})$ and f be defined as in* (3.3)*. Let $\mathbf{x}^*$ be an equilibrated local minimum of f for which Assumption* 1 *holds. Then the iteration*

$$\mathbf{x}^{(n+1)} = (R\circ \hat{S})(\mathbf{x}^{(n)})$$

is locally linearly convergent to $\mathbf{x}^*$.

3.4. Discussion of Assumption 1. We return to the least squares approximation. We wish to formulate conditions under which Assumption 1 will hold. This turns out to be quite subtle.

Define $\tau(\mathbf{x}) = \tau(\mathbf{A},\mathbf{B},\mathbf{C}) = \sum_{j=1}^{r} a_j \otimes b_j \otimes c_j$. Then $f(\mathbf{x}) = \frac{1}{2}\|\mathcal{T} - \tau(\mathbf{x})\|^2$ and

$$f''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}] = \|\tau'(\mathbf{x}^*)\mathbf{h}\|^2 + (\tau(\mathbf{x}^*) - \mathcal{T}, \tau''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}]). \tag{3.5}$$

Assumption 1 states that $f''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}] = 0$ only if $\mathbf{h} \in T\mathcal{M}^*_{\mathbf{x}^*}$. Note that for such $\mathbf{h}$ we necessarily have $\tau'(\mathbf{x}^*)\mathbf{h} = 0$ (since τ is constant on $\mathcal{M}^*$) and therefore also $(\tau(\mathbf{x}^*) - \mathcal{T}, \tau''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}]) = 0$.

Let us give an example, taken from [16], for which Assumption 1 does not hold. Consider $r = 3$ and $\mathcal{T}$ given pointwise by

$$\mathcal{T}_{i_1 i_2 i_3} = \sin(i_1 + i_2 + i_3).$$

One can prove that

$$\sin(i_1+i_2+i_3) = \sum_{j=1}^{3} \sin(i_j+\beta_j) \prod_{\substack{k=1\\k\neq j}}^{3} \frac{\sin(i_j+\beta_j+\alpha_k-\alpha_j)}{\sin(\alpha_k-\alpha_j)} \tag{3.6}$$

for all $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$ with $\sin(\alpha_k - \alpha_j) \neq 0$ for $j \neq k$ and all $\beta_1, \beta_2, \beta_3 \in \mathbb{R}$ with $\beta_1 + \beta_2 + \beta_3 = 0$. Hence, if $\tau(\mathbf{x}^*) = \mathcal{T}$ is any of the exact decompositions given by (3.6), the representation can be smoothly changed by manipulations beyond the scaling indeterminacy. Geometrically this means that τ and also f are constant on a submanifold of global minima of dimension more than $2r$. The null spaces of $\tau'(\mathbf{x}^*)$ and $f''(\mathbf{x}^*)$ are therefore larger than $T\mathcal{M}^*_{\mathbf{x}^*}$ and Assumption 1 cannot hold. A similar kind of example can be given for higher-order tensors.

Let us formulate the following assumption.

ASSUMPTION 2. *It holds that* $\tau'(\mathbf{x}^*)\mathbf{h} \neq 0$ *for all* $\mathbf{h} \notin T\mathcal{M}^*_{\mathbf{x}^*}$, *that is, the null space of* $\tau'(\mathbf{x}^*)$ *is* $T\mathcal{M}^*_{\mathbf{x}^*}$.

This assumption implies that the representation $\tau(\mathbf{x}^*)$ is locally essentially unique, which is necessary for Assumption 1, and excludes examples like (3.6). If now again $\tau(\mathbf{x}^*) = \mathcal{T}$ is an exact *decomposition*, then, by (3.5), Assumption 2 is even equivalent to Assumption 1.

THEOREM 3.5. *If* $\tau(\mathbf{x}^*) = \mathcal{T}$ *and Assumption* 2 *holds, then in a neighborhood of* $\mathbf{x}^*$ *the ALS algorithm is linearly convergent to* $\mathbf{x}^*$.

Let us now discuss the case of *approximation*, that is, the case $r < \operatorname{rank}\mathcal{T}$. Equation (3.5) suggests that Assumption 1 should hold if Assumption 2 holds and $\tau(\mathbf{x}^*) - \mathcal{T}$ is sufficiently small, that is, if $\tau(\mathbf{x}^*)$ is a good approximation for $\mathcal{T}$. More precisely, let W^* be any complementary space to $T\mathcal{M}^*_{\mathbf{x}^*}$, and let $\eta, \kappa > 0$ be such that

$$\|\tau'(\mathbf{x}^*)\mathbf{h}\|^2 \geq \eta\|\mathbf{h}\|^2, \quad \|\tau''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}]\| \leq \kappa\|\mathbf{h}\|^2 \tag{3.7}$$

for all $\mathbf{h} \in W^*$. Then, by (3.5), Assumption 1 will hold if

$$\|\mathcal{T} - \tau(\mathbf{x}^*)\| < \frac{\eta}{\kappa}. \tag{3.8}$$

However, to obtain a reasonable a priori statement of how close $\tau(\mathbf{x}^*)$ has to be to $\mathcal{T}$, one has to express η and κ in terms of $\mathcal{T}$ only. Or, since $\mathcal{T}$ and $\tau(\mathbf{x}^*)$ are related via

$$\|\tau(\mathbf{x}^*)\|^2 = \|\mathcal{T}\|^2 - \|\mathcal{T} - \tau(\mathbf{x}^*)\|^2$$

(which is just the necessary optimality condition, or normal equation, for (2.2) on the span of $\tau(\mathbf{x}^*)$), one may equivalently try to find constants η and κ which depend on $\tau(\mathbf{x}^*)$ only. The problem is that this is usually not possible for κ.[6] As so often is the case, we will have to content ourselves with an investigation of the rank-one approximation problem.

3.4.1. A sufficient condition for the rank-one case. If $r = 1$, we have to show $\operatorname{rank} f''(\mathbf{x}^*) = n_1 + n_2 + n_3 - 2$. Since $\mathbf{x}^* \in \hat{\mathcal{X}}$, a space of this dimension is given by

$$W^* = \{\mathbf{h} = (\delta a, \delta b, \delta c) \in \mathcal{X} \mid \delta a \perp a^*,\ \delta b \perp b^*\}.$$

We have $\tau(\mathbf{x}^*) = a^* \otimes b^* \otimes c^*$,

$$\tau'(\mathbf{x}^*)\mathbf{h} = \delta a \otimes b^* \otimes c^* + a^* \otimes \delta b \otimes c^* + a^* \otimes b^* \otimes \delta c$$

[6] As far as we can see, this would amount to estimating $\sum_{j=1}^r \|a_j^* \otimes b_j^* \otimes c_j^*\|$ in terms of $\tau(\mathbf{x}^*)$, which for $r \geq 2$ cannot be achieved without further information on the representation $\mathbf{x}^*$. This is closely related to the phenomenon that the canonical low-rank approximation can be an ill-posed problem ("diverging rank-one terms").

and

$$\tau''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}] = 2(\delta a \otimes \delta b \otimes c^* + \delta a \otimes b^* \otimes \delta c + a^* \otimes \delta b \otimes \delta c).$$

Let $\sigma = \|\tau(\mathbf{x}^*)\|$, that is, $\sigma^{1/3} = \|a^*\| = \|b^*\| = \|c^*\|$. Using (1.1), we see that for $\mathbf{h} \in W^*$ the terms in $\tau'(\mathbf{x}^*)$ are pairwise orthogonal so that

$$\|\tau'(\mathbf{x}^*)\mathbf{h}\|^2 = \sigma^{4/3}\|\mathbf{h}\|^2.$$

Hence we can choose $\eta = \sigma^{4/3}$ in (3.7). In particular, Assumption 2 is always satisfied in the rank-one case.

To estimate $\|\tau''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}]\|$, let $\|\mathbf{h}\| = 1$. Using that for $\mathbf{h} \in W^*$ the terms in $\tau''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}]$ are also pairwise orthogonal, one may verify that, under the constraint $\|\mathbf{h}\| = 1$, the norm $\|\tau''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}]\|^2$ is maximal only if $\|\delta a\| = \|\delta b\| = \|\delta c\| = 1/\sqrt{3}$. This gives

$$\|\tau''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}]\| \leq \kappa = 2\sqrt{3 \cdot \Big(\frac{\sigma^{1/3}}{3}\Big)^2} = \frac{2}{\sqrt{3}}\sigma^{1/3}.$$[7]

We conclude from (3.8) that Assumption 1 is valid if

$$\|\mathcal{T} - \tau(\mathbf{x}^*)\| < \frac{\eta}{\kappa} = \frac{\sqrt{3}}{2}\sigma,$$

or, using $\sigma^2 = \|\mathcal{T}\|^2 - \|\mathcal{T} - \tau(\mathbf{x}^*)\|^2$ and rearranging,

$$\|\mathcal{T} - \tau(\mathbf{x}^*)\| < \sqrt{\frac{3}{7}}\|\mathcal{T}\|.$$

For higher-order $d > 3$ the same reasoning leads to

$$\eta = \sigma^{2(d-1)/d}$$

and

$$\kappa = 2\sqrt{\frac{d(d-1)}{2} \cdot \Big(\frac{\sigma^{(d-2)/d}}{d}\Big)^2} = \sqrt{\frac{2d-2}{d}}\sigma^{(d-2)/d}$$

(since $\tau''(\mathbf{x}^*)[\mathbf{h},\mathbf{h}]$ consists of $d(d-1)/2$ pairwise orthogonal terms then), so that (3.8) becomes

$$\|\mathcal{T} - \tau(\mathbf{x}^*)\| < \sqrt{\frac{d}{2d-2}}\sigma,$$

or, equivalently,

$$\|\mathcal{T} - \tau(\mathbf{x}^*)\| < \sqrt{\frac{d}{3d-2}}\|\mathcal{T}\|.$$

If $\tau(\mathbf{x}^*)$ is supposed to be the *best* rank-one approximation to $\mathcal{T} \neq 0$ (which always exists), we can formulate the following criterion.

[7] This, by the way, is the crucial estimate which fails for matrices ($d = 2$), since in that case $\tau(\mathbf{x}^*)''[\mathbf{h},\mathbf{h}] = 2\ \delta a \otimes \delta b$ does not depend on $\mathbf{x}^*$ anymore.

THEOREM 3.6. *Let $d \geq 3$. If the Euclidian distance between an order-d tensor $\mathcal{T}$ and the set of rank-one tensors is strictly smaller than $\|\mathcal{T}\|/\sqrt{3-2/d}$, then Assumption* 1 *holds for any best rank-one approximation of $\mathcal{T}$. Consequently, the ALS algorithm then converges linearly in a neighborhood of a best rank-one approximation.*

Although this is a surprisingly soft condition, we remark that it may not be necessary for Assumption 1 to hold. In particular, condition (3.8) might be too conservative, since it stems from the Cauchy–Schwarz inequality.

The matrix case $d = 2$ is not covered by the above estimate. Even though the given proof already fails (see footnote 7), let us consider an example. Let $\mathcal{T}$ be the 2×2 identity matrix, which is the worst case when it comes to rank-one approximation. Its distance to the set of rank-one matrices is 1, which is smaller than $\|\mathcal{T}\|/\sqrt{3-2/d} = \sqrt{2}$. But every matrix

$$\begin{pmatrix} \sin^2 t & \sin t \cos t \\ \sin t \cos t & \cos^2 t \end{pmatrix} = \begin{pmatrix} \sin t \\ \cos t \end{pmatrix} \otimes \begin{pmatrix} \sin t \\ \cos t \end{pmatrix} \tag{3.9}$$

is a best rank-one approximation of the identity. Obviously, this set cannot be obtained by rescaling a single rank-one matrix. Therefore this again is an example where Assumption 1 does not hold.[8]

The rank-one case has also been studied in [27], but within a different framework. A quite general condition has been formulated there for the local convergence of ALS, namely the positive definiteness of a certain Lagrangian. As we believe, it is closely related to our Assumption 1.

3.5. A note on regularization. There are several reasons why one should consider instead of (2.2) a Tikhonov regularized problem such as

$$g_\lambda(\mathbf{A}, \mathbf{B}, \mathbf{C}) = f(\mathbf{A}, \mathbf{B}, \mathbf{C}) + \lambda(\|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 + \|\mathbf{C}\|^2) = \min, \tag{3.10}$$

where $\lambda > 0$ is a regularization parameter and the norms are the corresponding Euclidian (Frobenius) matrix norms. (Note that $\|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 + \|\mathbf{C}\|^2 = \|\mathbf{x}\|^2$.)

The first reason is that this problem is always well-posed [13], that is, admits a global minimizer (it is coercive). This also has direct consequences to the behavior of an ALS algorithm applied to (3.10). It has been observed that swamps occur less often so that (global) convergence to a critical point is generally faster [17, 21].

A second, equally important reason to consider (3.10) is that the scaling indeterminacy is completely removed. One can check that for a local minimum $\mathbf{x}^* \in \hat{\mathcal{X}}$ of (3.10) it necessarily holds that $\|a_j^*\| = \|b_j^*\| = \|c_j^*\|$ for all $j = 1, 2, \ldots, r$, that is, the solution is equilibrated.[9] It is hence reasonable to apply the standard convergence theory of the nonlinear block Gauss–Seidel method by assuming that $g_\lambda''(\mathbf{x}^*)$ is positive definite at a local solution. Since Lemma 3.2 still holds, the local convergence of ALS applied to g_λ (without the equilibration R) then follows immediately from the above considerations; cf. [1, Theorem 2] and [18, 10.3.5]. Indeed, we have the following theorem.

THEOREM 3.7. *If λ is large enough, $g_\lambda''(\mathbf{x}^*)$ is positive definite at* global *minimizers $\mathbf{x}^*$. Consequently, the ALS algorithm (without equilibration) is locally linearly convergent at such points.*

[8] For higher ranks, Assumption 1 will never hold in the matrix case, since a low-rank decomposition UV^T might be replaced by $UAA^{-1}V^T$, which is more than just a scaling indeterminacy.

[9] Using the Lagrange multiplier rule, one can show that among all rescalings $\alpha_j a_j^*$, $\beta_j b_j^*$, $\gamma_j c_j^*$ with $\alpha_j \beta_j \gamma_j = 1$ only the one that leads to an equilibrated solution minimizes the second term in (3.10).

Proof. Fix $\lambda_0 > 0$. All global minima $\mathbf{x}^*_{\lambda_0}$ of g_{λ_0} lie in a certain ball of radius depending on λ_0 (again since g_{λ_0} is coercive). Then for $\lambda > \lambda_0$ the global minima $\mathbf{x}^*_\lambda$ of g_λ also have to lie in that ball, since otherwise we would obtain the contradiction

$$g_\lambda(\mathbf{x}^*_{\lambda_0}) = g_{\lambda_0}(\mathbf{x}^*_{\lambda_0}) + (\lambda - \lambda_0)\|\mathbf{x}^*_{\lambda_0}\| < g_{\lambda_0}(\mathbf{x}^*_\lambda) + (\lambda - \lambda_0)\|\mathbf{x}^*_\lambda\| = g_\lambda(\mathbf{x}^*_\lambda).$$

Consequently, the Hessian $f''(\mathbf{x}^*_\lambda)$ can be bounded by a constant depending only on λ_0, so that

$$g''_\lambda(\mathbf{x}^*_\lambda) = f''(\mathbf{x}^*_\lambda) + 2\lambda I$$

will be positive definite if λ is large enough. ∎

Of course one does not want to make λ too large. At least, in contrast to the unregularized case, for every λ the heuristic mentioned in the previous section can be justified for global minima: since $\tau''(\mathbf{x}^*_\lambda)$ now only has to be bounded on the ball in which the global minimizers $\mathbf{x}^*_\lambda$ are located (which depends only on λ), the Hessian

$$g''_\lambda(\mathbf{x}^*_\lambda)[\mathbf{h},\mathbf{h}] = \|\tau'(\mathbf{x}^*_\lambda)\mathbf{h}\|^2 + (\tau(\mathbf{x}^*_\lambda) - \mathcal{T}, \tau''(\mathbf{x}^*_\lambda)[\mathbf{h},\mathbf{h}]) + 2\lambda\|\mathbf{h}\|^2$$

will be positive definite if $\tau(\mathbf{x}^*_\lambda) - \mathcal{T}$ is smaller than a certain constant depending on $\mathcal{T}$ and λ only. In particular, for exact decompositions the Hessian is always positive definite, even if Assumption 2 is not satisfied.

4. Numerical experiments. Our numerical experiments were quite simple and only meant to demonstrate the linear convergence rate and check the size of the convergence region. The calculations have been neither systematic nor exhaustive. Real applications can be found elsewhere. We emphasize again that we were interested in the local convergence of the factor matrices, that is, say in the third-order case, we measured the Euclidian norm

$$\|\mathbf{x} - \mathbf{x}^*\| = \sqrt{\|\mathbf{A} - \mathbf{A}^*\|^2 + \|\mathbf{B} - \mathbf{B}^*\|^2 + \|\mathbf{C} - \mathbf{C}^*\|^2},$$

where the matrix norms are the Frobenius norms. To obtain a relative measure, we normalized the columns of the factor matrices to one (instead of equilibrating them). Note that the norms $\|\tau(\mathbf{x}) - \tau(\mathbf{x}^*)\|$, due to the local Lipschitz continuity of τ, would produce almost the same semilogarithmic plot (cf. Theorem 3.3), but give no information about the convergence of the factor matrices.

In our experiments we randomly generated tensors of different size and order and first calculated a "best" rank-r approximation using ALS with a random starting point. As a stopping criterion we imposed that the difference $\|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\|$ between two subsequent normalized iterates should be sufficiently small ($\sim 10^{-14}$) in order to "guarantee" that the solution is at least a critical point of the approximation problem. For large values of r we had to try many starting points to achieve this precision due to the appearance of swamps.

The normalized factor matrices of the calculated solutions were randomly perturbed by different orders of magnitude and then used as a starting point for a restart of the ALS algorithm. We observed that the initial solution would be recovered if the perturbation was of magnitude at most 10^0. In fact, larger perturbations result in an almost random starting point, so the ALS iteration almost surely will converge to a different critical point. Usually, the convergence rate decreased for larger r, but not in all experiments. In Figure 4.1 we plotted the errors $\|\mathbf{x}^* - \mathbf{x}^{(n)}\|$ for the rank $r = 2, 3, 4$, and 5 approximations of a random, but fixed, $10 \times 10 \times 10$ tensor. As one

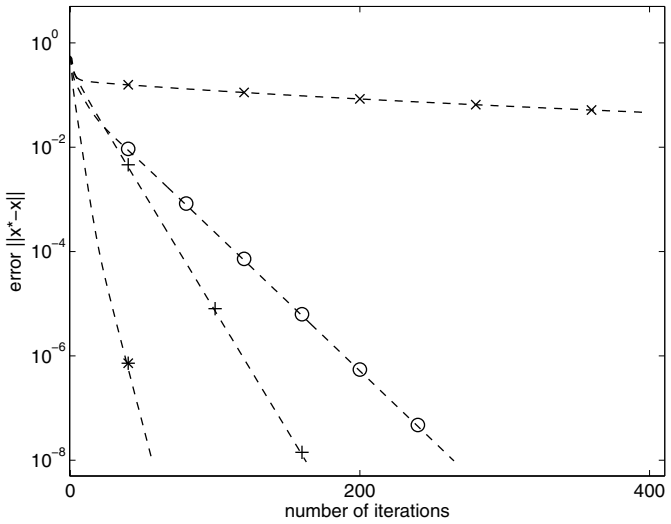

FIG. 4.1. *Depicted are the absolute errors* $\|\mathbf{x}^* - \mathbf{x}^{(n)}\|$ *between "exact" normalized factor matrices of a rank-r approximation of a random* $10 \times 10 \times 10$ *tensor and the iterates of the corresponding ALS algorithm. Plots are given for* $r = 2$ $(*)$, $r = 3$ $(\circ)$, $r = 4$ $(+)$, *and* $r = 5$ $(\times)$. *As one can see, the convergence rate is better for* $r = 4$ *than for* $r = 3$. *However, this behavior was rather exceptional in our experiments.*

can see, the convergence is linear and the rate is better for $r = 4$ than for $r = 3$ in this example, but then again really slow for $r = 5$.

For special tensors, such as hyperdiagonal or, more generally, complete orthogonal tensors [10], one can exactly determine the best rank-r approximation by truncation of a complete orthogonal representation. Applying the same kind of experiment for such tensors we observed very fast convergence of ALS (at most four iterations independent of r), so we did not include plots for this case.

5. Summary. The easiest way to prove local linear convergence of a nonlinear fixed-point iteration consists of showing a contraction property of its linearization at the fixed point. In the case of the naive ALS method (2.3), the linearization is the block Gauss–Seidel method for the Hessian $f''(\mathbf{x}^*)$ of the target functional at the fixed point, which fails to be contractive if $f''(\mathbf{x}^*)$ is not positive definite. Unfortunately, this happens for the considered target functional (2.1) due to the scaling indeterminacy of the canonical tensor format: the Hessian $f''(\mathbf{x}^*)$ possesses a null space which at least contains the tangent space (3.1) to the orbit of rescaled representations of $\sum_{j=1}^{r} a_j^* \otimes b_j^* \otimes c_j^*$.

The idea of our convergence proof was to realize that a suitably chosen equilibration operator, which fixes the canonical representation and is employed in practice to stabilize the iteration, removes the parts belonging to this tangent space from the first-order terms of the error iteration (Lemma 3.1). After assuming that the Hessian is positive in all other directions, local convergence of Algorithm 1 was routinely established (Theorem 3.3).

This main Assumption 1 in the convergence theorem is closely related to the question of (local) essential uniqueness of the CP representation of the solution (this being a necessary condition). Therefore, it will not always hold (example (3.6)) and might be considered controversial. At least, we were able to derive a condition which guarantees

its validity for global minima of the rank-one approximation (Theorem 3.6). In any case, we think that the assumption cannot be avoided when aiming at a local convergence proof of ALS in the stated generality via the contraction principle.

REFERENCES

[1] J.C. Bezdek and R.J. Hathaway, *Convergence of alternating optimization*, Neural Parallel Sci. Comput., 11 (2003), pp. 351–368.

[2] G. Beylkin and M.J. Mohlenkamp, *Algorithms for numerical analysis in high dimensions*, SIAM J. Sci. Comput., 26 (2005), pp. 2133–2159.

[3] 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.

[4] 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.

[5] W.H. Greub, *Multilinear Algebra*, Springer-Verlag, New York, 1967.

[6] L. Grippo and M. Sciandrone, *Globally convergent block-coordinate techniques for unconstrained optimization*, Optim. Methods Softw., 10 (1999), pp. 587–637.

[7] 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.

[8] R.A. Harshman, *Determination and proof of minimum uniqueness conditions for PARAFAC*1, UCLA Working Papers in Phonetics, 22 (1972), pp. 111–117.

[9] H.B. Keller, *On the solution of singular and semidefinite linear systems by iteration*, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal., 2 (1965), pp. 281–290.

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

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

[12] 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.

[13] L.-H. Lim and P. Comon, *Nonnegative approximations of nonnegative tensors*, J. Chemometrics, 23 (2009), pp. 432–441.

[14] 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.

[15] M.J. Mohlenkamp, *Musings on multilinear fitting*, Linear Algebra Appl., in press.

[16] M.J. Mohlenkamp and L. Monzón, *Trigonometric identities and sums of separable functions*, Math. Intelligencer, 27 (2005), pp. 65–69.

[17] C. Navasca, L.D. Lathauwer, and S. Kindermann, *Swamp reducing technique for tensor decomposition*, in Proceedings of the 16th European Signal Processing Conference, Lausanne, 2008.

[18] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970. Reprinted as Classics Appl. Math. 30, SIAM, Philadelphia, 2000.

[19] I.V. Oseledets, *On a new tensor decomposition*, Dokl. Math., 80 (2009), pp. 495–496; translated from Dokl. Akad. Nauk, 427 (2009), pp. 168–169.

[20] I.V. Oseledets, *Tensor-train decomposition*, SIAM J. Sci. Comput., 33 (2011), pp. 2295–2317.

[21] P. Paatero, *Construction and analysis of degenerate PARAFAC models*, J. Chemometrics, 14 (2000), pp. 285–299.

[22] M. Rajih, P. Comon, and R.A. Harshman, *Enhanced line search: A novel method to accelerate PARAFAC*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1128–1147.

[23] S. Schechter, *Iteration methods for nonlinear problems*, Trans. Amer. Math. Soc., 104 (1962), pp. 179–189.

[24] J.M.F. Ten Berge and J.N. Tendeiro, *The link between sufficient conditions by Harshman and by Kruskal for uniqueness in Candecomp/Parafac*, J. Chemometrics, 33 (2009), pp. 321–323.

[25] G. Tomasi and R. Bro, *A comparison of algorithms for fitting the PARAFAC model*, Comput. Statist. Data Anal., 50 (2006), pp. 1700–1734.

[26] J. Wu, Y.-J. Lee, J. Xu, and L. Zikatanov, *Convergence analysis on iterative methods for semidefinite systems*, J. Comput. Math., 26 (2008), pp. 797–815.

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