SIAM J. NUMER. ANAL.
Vol. 51, No. 2, pp. 1134–1162

ON LOCAL CONVERGENCE OF ALTERNATING SCHEMES FOR OPTIMIZATION OF CONVEX PROBLEMS IN THE TENSOR TRAIN FORMAT*

THORSTEN ROHWEDDER† AND ANDRÉ USCHMAJEW†

Abstract. Alternating linear schemes (ALS), with the alternating least squares algorithm a notable special case, provide one of the simplest and most popular choices for the treatment of optimization tasks by tensor methods. An according adaptation of ALS for the recent TT (= tensor train) format [I. V. Oseledets, *SIAM J. Sci. Comput.*, 33 (2011), pp. 2295–2317], known in quantum computations as matrix product states, has recently been investigated in [S. Holtz, T. Rohwedder, and R. Schneider, *SIAM J. Sci. Comput.*, 34 (2012), pp. A683–A713]. With the present work, the positive practical experience with TT-ALS is backed up with an according local linear convergence theory for the optimization of convex functionals J. The main assumption entering the proof is that the redundancy introduced by the TT parametrization τ matches the null space of the Hessian of the induced functional $j = J \circ \tau$, and we give conditions under which this assumption can be expected to hold. In particular, this is the case if the TT rank has been correctly estimated. The case of nonconvex functionals J is also shortly discussed.

Key words. ALS, high-dimensional optimization, local convergence, matrix product states, nonlinear Gauss–Seidel, tensor product approximation, tensor train decomposition

AMS subject classifications. 15A69, 65K10, 90C06

DOI. 10.1137/110857520

1. Introduction.

In many application areas, the treatment of the respective governing equations amounts to the treatment of discrete tensors, i.e., of high dimensional quantities $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$. For example, such problems arise in the context of life sciences and physics in the discretization of functions from tensor spaces, often defined implicitly as the solution of e.g., partial differential or integral equations, with various Schrödinger equation type models and the Fokker–Planck equation providing prominent examples. Another field where problems posed on high-dimensional spaces turn up naturally is the active field of data mining problems. Since the standard approaches to all such problems have a computational complexity growing exponentially in the dimension d of the tensors, the only effective remedy is often the use of a data-sparse representation or approximation of the tensors exploiting concepts of tensor product approximations, and the development of such concepts has consequently become an important and active field of mathematical research during the last years. An exhaustive overview over the current state of the subject is provided by Hackbusch's recently issued book [16].

As a classical approach to multidimensional problems, the canonical format (also known as the CANDECOMP/PARAFAC model, [27]) is extensively used in practical applications for extraction of information, i.e., linear least squares problems. This is contrasted by a lack of desirable theoretical properties [6]. Also, for treatment of more complex equations as optimization problems, application of this format is possible [11], but has to be stabilized by somewhat artificial techniques [9]; when it comes

*Received by the editors December 1, 2011; accepted for publication (in revised form) December 26, 2012; published electronically April 4, 2013.
http://www.siam.org/journals/sinum/51-2/85752.html
†Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany (Current e-mail: rohwedde@math.hu-berlin.de, andre.uschmajew@epfl.ch). This research was supported by DFG (Deutsche Forschungsgemeinschaft) through SPP 1324.

to the treatment of equations based on a manifold approach [18], CANDECOMP lacks the basic property of being an embedded manifold, meaning that a stable, nonredundant parametrization of the set of rank-r tensors cannot be given, ruling out the canonical format for such approaches. An alternative is provided by the—also quite classical—Tucker format [41], with the set of Tucker tensors of bounded rank forming a weakly closed set [12] with manifold properties [26], fostering, for instance, its application in quantum chemistry [4]. In practical application of the Tucker format, one is then unfortunately confronted with the problem that although it sometimes reduces complexity immensely, the ansatz still scales exponentially in d, often prohibiting its application to problems of higher dimension d. Hackbusch, therefore, generalized the idea of subspace approximation, being the basis of the Tucker decomposition, to a hierarchical ansatz proposed in [17]: High-dimensional tensors are therein successively decomposed to component tensors of order 3, which then can be treated separately, so that the complexity stays linear or at least polynomial in the dimension d. Many desirable theoretical features of this approach have recently been verified: Results on the existence of best approximations and minimizers of convex functionals from [12, 13] also apply to the hierarchical Tucker format; extending results from [20], manifold properties can be verified [43].

From the practical side, two notable special cases of the hierarchical ansatz have received major attention: The HT format as used by Grasedyck [14], using a balanced decomposition tree, and the tensor train (TT) format, as introduced independently by Oseledets [33] and Tyrtyshnikov and Oseledets [37], using a linear decomposition tree. For both approaches, the applicability to various problems has lately been investigated with promising results (see, e.g., [1, 2, 21, 25, 28]), while the former approach seems to be superior on some problems in the sense that it sometimes provides lower ranks [15], the latter has the advantage of being conceptually quite simple, thus often allowing for a simpler analysis from the theoretical point of view, and enabling, from the practical side, a robust sequential treatment of resulting equations to be solved for the component tensors. In particular, one of the authors of the present publication was co-author of the previous work [21], in which the well-known alternating least squares has been adapted to the treatment of more general high-dimensional optimization tasks like linear equations, eigenvalue equations etc. using the TT format. We found that stable equations can be derived from the concept of left- and right-orthogonality; proceeding in "sweeps" inspired by the DMRG algorithm used in quantum chemistry [39, 48, 47] gives an iteration scheme that we termed the alternating linear scheme for optimization in the TT format (abbreviated by ALS in what follows, introduced alongside with a modified variant MALS enabling dynamical rank-adaptation like in DMRG; see [21]). In this context we should note that the TT format is known in quantum computations under the name *matrix product states* (MPS) for many years. This name is perfectly justified by the definition (1.1). However, the name *tensor train* also has its figurative meaning and is now quite common in the mathematical community, so that, in accordance with previous papers, we have chosen it here again. One will have to judge on this decision in the future. We refer to the recent survey [40] by Schollwöck on MPS and DMRG for the physical point of view and historical notes.

Our practical experience, as partly reported in [21], is that the ALS shows extraordinarily nice convergence behavior similar to that of ALS applied to the Tucker format. Although other approaches based on tangent space of the TT manifold may be more suitable for treatment of e.g., differential equations [18, 20, 30, 43], and although ALS might be refined by using locally quadratically convergent methods near

the sought optimum, the ALS approach is, in fact, at the moment our basic method of choice for the treatment of optimization problems in the sense that even with its striking simplicity, it converges stably, efficiently, and reliably to stationary points of the parametrized problem. Of course, these points might not be the global minima on the approximation manifold under consideration—it has recently even been shown that the computation of global rank-one-best approximations is an NP-hard problem [19]. The occurrence of local minima is a (not solely theoretical) problem common to many tensor optimization methods; in the case of ALS for the TT format, it is often treated by random perturbation of computed stationary points (e.g., in practical applications from quantum chemistry), and has for the DMRG/MALS algorithm also readily been attacked in the recent publication [35] in a more systematic way.

The aim of the present paper is to back-up our positive practical experiences with ALS with an according theoretical investigation of the convergence behavior of the algorithm when applied in the above form to the treatment of convex optimization problems in the TT format. Our practical observation in [21] was that ALS provides a very robust linear convergence behavior. In this work we deliver the according theoretical result, by rigorously proving local linear convergence under certain reasonable positive definiteness conditions on the Hessian of the functional to be minimized. Our proof is an adaptation of the concept followed in the recent paper [42], where one of the authors investigated local convergence of PARAFAC-ALS: in the neighborhood of a solution, the TT-ALS algorithm is identified as a perturbation of the linear block Gauss–Seidel method applied to the Hessian at the solution [31, 32]. This Hessian is only semidefinite due to the nonuniqueness of the TT representation. In contrast to the linear case, one has to completely remove the null space of the Hessian from the iteration to keep the contractive properties of the linear Gauss–Seidel method [46, 24, 29]. This can be achieved by introducing a local normalization operator which chooses a unique representation for the TT tensors. While this is technically necessary to link the convergence analysis to existing results on the nonlinear Gauss–Seidel method for semidefinite problems, it turns out that the sequence of TT tensors generated by the ALS algorithm is independent of the choice of their representations therein (Proposition 2.3). From the convergence result for *one* choice of representations one hence obtains a practically relevant convergence result for a whole equivalence class of *all* possible representations. In other words, a convergence result for ALS regarded as an algorithm on the manifold of tensors of fixed TT rank.

In short, the content of this paper can be outlined as follows. In section 2, we devise a generic ALS algorithm for convex functionals for which we prove convergence under assumptions on the solution and on the Hessian of the functional (Theorem 2.7 and Corollaries 2.8 and 2.9). Based on this, convergence for the ALS from [21] is deduced (Theorem 2.10). In section 2.5 it is explained why the convexity assumption can be dropped in a local procedure. Sufficient conditions for the main assumption on the Hessian to hold in the convex case are presented in sections 2.4 and 3.

The TT format is introduced in nearly every paper on the subject. Below we just give a very short definition. We added an appendix which contains a short primer on the TT rank and the well-known problem of nonunique representations which plays the crucial role in the convergence analysis. We discuss this issue from a quite geometric viewpoint, which will be fully developed in the upcoming publication [43].

1.1. The TT decomposition. Let $d \in \mathbb{N}$ and $n_1, n_2, \dots, n_d \in \mathbb{N}$ be given. We use the notation $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ for the space of dth order tensors and treat its elements

$\mathbf{X}$ as d-dimensional arrays, whose entries are indexed by

$$\mathbf{X}(x_1, x_2, \ldots, x_d), \qquad x_i = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, d.$$

Let $\mathbf{r} = (r_1, r_2, \ldots, r_{d-1})$, where $0 < r_i \leq n_i$ are integers. Further we set $r_0 = r_d = 1$. The elements of the space

$$\overline{\mathcal{U}} = \bigtimes_{i=1}^{d} \mathbb{R}^{r_{i-1} \times n_i \times r_i}$$

are denoted by $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_d)$. The x_ith lateral slice of a third-order component tensor $\mathbf{U}_i \in \mathbb{R}^{r_{i-1} \times n_i \times r_i}$ will be denoted by $U_i(x_i)$, that is, for x_i fixed $U_i(x_i)$ is the $r_{i-1} \times r_i$ matrix given by

$$[U_i(x_i)]_{k_1 k_2} = \mathbf{U}_i(k_1, x_i, k_2).$$

Consider the mapping

$$\tau \colon \overline{\mathcal{U}} \to \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \colon \mathbf{U} \mapsto \tau(\mathbf{U}) = \tau(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_d)$$

given pointwise by

$$\tau(\mathbf{U})(x_1, x_2, \ldots, x_d) := U_1(x_1) U_2(x_2) \ldots U_d(x_d). \tag{1.1}$$

Since $r_0 = r_d = 1$, this matrix product indeed results in a real number for each multi-index $(x_1, x_2, \ldots, x_d)$.

DEFINITION 1.1 ($\mathrm{TT}_{\leq \mathbf{r}}$ format). *A tensor of the form $\mathbf{X} = \tau(\mathbf{U})$ is called a* $\mathrm{TT}_{\leq \mathbf{r}}$ tensor *or a* tensor in $\mathrm{TT}_{\leq \mathbf{r}}$ format*, and $\mathbf{U}$ is called a* $\mathrm{TT}_{\leq \mathbf{r}}$ decomposition *of $\mathbf{X}$. The image of τ is denoted by*

$$\mathcal{T}_{\leq \mathbf{r}} = \tau(\overline{\mathcal{U}}) \subseteq \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$$

and is called the set of TT tensors of rank at most $\mathbf{r}$.

The mapping $(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_d) \mapsto \tau(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_d)$ is multilinear and homogenous, so that the $\mathrm{TT}_{\leq \mathbf{r}}$ decomposition has a lot of nice structural properties which are similar to those of rank-one tensors, and thus allow for an analogous treatment, at least to some extent. In fact, rank-one tensors are nonzero $\mathrm{TT}_{\leq \mathbf{r}}$ tensors where all $r_i = 1$.

In what follows we will mainly focus on the set $\mathcal{T}_{\mathbf{r}}$ of tensors of TT rank *equal* to $\mathbf{r}$. The definition of the rank is based on the left and right unfoldings of the components $\mathbf{U}_i$, which will be denoted by $\mathbf{U}_i^L \in \mathbb{R}^{r_{i-1} n_i \times r_i}$ and $\mathbf{U}_i^R \in \mathbb{R}^{r_{i-1} \times n_i r_i}$. In [21] the notation $L(\mathbf{U}_i)$ and $R(\mathbf{U}_i)$ has been used, respectively. According to Proposition A.2, $\mathcal{T}_{\mathbf{r}}$ is parametrized by the open and dense subset

$$\mathcal{U} = \{\mathbf{U} \in \overline{\mathcal{U}} \mid \operatorname{rank} \mathbf{U}_i^L = \operatorname{rank} \mathbf{U}_{i+1}^R = r_i \text{ for } i = 1, 2, \ldots, d-1\}$$

of $\overline{\mathcal{U}}$, that is, open and dense in $\overline{\mathcal{U}}$, that is, we have $\mathcal{T}_{\mathbf{r}} = \tau(\mathcal{U})$. Here it is silently assumed that $\mathcal{U}$ is not empty. For details we refer to Appendix A.1.

Finally, let

$$\mathsf{A} = (A_1, A_2, \ldots, A_{d-1}) \in \mathcal{G} = \bigtimes_{i=1}^{d-1} GL(r_i)$$

be a tuple of invertible matrices $A_i \in \mathbb{R}^{r_i \times r_i}$. By looking at (1.1), we have $\tau(\mathbf{U}) = \tau(\hat{\mathbf{U}})$, where

$$\hat{U}_1(x_1) = U_1(x_1)A_1, \quad \hat{U}_d(x_d) = A_{d-1}^{-1}U_d(x_d), \quad \text{and} \quad \hat{U}_i(x_i) = A_{i-1}^{-1}U_i(x_i)A_i. \tag{1.2}$$

This operation will be called a *rescaling by* A, and denoted by $\hat{\mathbf{U}} = \theta_{\mathbf{U}}(\mathsf{A})$. This map can be regarded as a Lie group action on the paramter set $\overline{\mathcal{U}}$ which generates equivalent $\mathrm{TT}_{\leq \mathbf{r}}$ parametrizations

$$\mathcal{M}_{\mathbf{U}} = \{\hat{\mathbf{U}} = \theta_{\mathbf{U}}(\mathsf{A}) \colon \mathsf{A} \in \mathcal{G}\}$$

of the same tensor $\tau(\mathbf{U})$. Hence the $\mathrm{TT}_{\leq \mathbf{r}}$ decomposition is not unique. This is discussed in detail in Appendices A.2 and A.3. The set $\mathcal{M}_{\mathbf{U}}$ will be called the *orbit* of $\mathbf{U}$.

2. Local convergence of the ALS. In this section we present an abstract alternating optimization scheme which involves rescalings during the iteration process. Taking rescaling into account is important since the stability of practical computations in the TT format partially relies on some sort of repeated orthogonalization procedure during the process. As it turns out, surprisingly, changing the $\mathrm{TT}_{\leq \mathbf{r}}$ representation between alternating optimization of the single components $\mathbf{U}_i$ does not make the analysis more difficult, but is even helpful.

First, local convergence of a particular fixed point version of the algorithm (using local normal forms) will be proved under the assumption that the Hessian of the loss function j to be defined below (see (2.2)) is positive definite at the solution modulo the null space caused by the nonuniqueness (1.2) of the $\mathrm{TT}_{\leq \mathbf{r}}$ representation. Realizing that convergence in the sense of orbits is not affected when the iterates are moved on their orbits during the iteration, we will be able to deduce a convergence result for the general version of the algorithm, and, as a special case, for the alternating linear scheme introduced in [21]. We also give a short treatise on ALS for nonconvex functionals and on its application to reconstruction problems. The discussion of the positive definiteness assumption is postponed to the next section.

2.1. Alternating optimization and scaling. Let

$$J \colon \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}$$

be a *strictly convex* C^2-functional to be minimized. *We assume that J possesses a minimizer on* $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, which implies that J is *coercive*. Due to the high-dimensionality of the domain, the task is restricted to the set of $\mathrm{TT}_{\leq \mathbf{r}}$ decomposable tensors:

$$J(\mathbf{X}) = \min, \quad \mathbf{X} \in \mathcal{T}_{\leq \mathbf{r}}. \tag{2.1}$$

In the parametrized version, let

$$j = J \circ \tau \colon \overline{\mathcal{U}} \to \mathbb{R} \colon \mathbf{U} \mapsto J(\tau(\mathbf{U})); \tag{2.2}$$

then we seek for a solution of

$$j(\mathbf{U}) = j(\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_d) = \min. \tag{2.3}$$

A prominent example is the problem of finding a best approximation of a tensor $\mathbf{Y}$ by a $\mathrm{TT}_{\leq \mathbf{r}}$ decomposable tensor in the Frobenius (= Euclidian) norm, such that the

task becomes $\|\mathbf{Y}-\tau(\mathbf{U})\|_F^2 = \min$. Another important application is the approximate solution of linear systems in TT format, $\|\mathbf{A}\tau(\mathbf{U}) - b\| = \min$, where $\mathbf{A}$ is ideally a tensor structured invertible or even positive definite linear operator [21, 35].

Obviously, if $\mathbf{X}^* = \tau(\mathbf{U}^*)$ is a local minimum (2.1) (local in $\mathcal{T}_{\leq \mathbf{r}}$), then $\mathbf{U}^*$ has to be a local minimum of j in $\overline{\mathcal{U}}$. The converse is at least true if $\mathbf{U}^* \in \mathcal{U}$. Namely, since submersions are open maps, it follows from Proposition A.3 that $\mathbf{X}^*$ is a local minimum of J in $\mathcal{T}_{\mathbf{r}}$. Since this set is open in $\mathcal{T}_{\leq \mathbf{r}}$ (for $\mathcal{T}_{\leq \tilde{\mathbf{r}}}$ is closed for every $\tilde{\mathbf{r}} \leq \mathbf{r}$), $\mathbf{X}^*$ is also a local minimum in $\mathcal{T}_{\leq \mathbf{r}}$.

Our results will hold only for (local) solutions $\mathbf{U}^* \in \mathcal{U}$ of (2.3), that is, for (local) solutions $\mathbf{X}^* \in \mathcal{T}_{\mathbf{r}}$. We assume such solutions to exist. In particular, $\mathcal{U}$ shall not be empty.

Note that j is constant on orbits $\mathcal{M}_{\mathbf{U}} \subseteq \mathcal{U}$, which has the following consequence.

PROPOSITION 2.1. *Let $\mathbf{U}^* \in \mathcal{U}$ be a local minimizer of j. Then every $\hat{\mathbf{U}}^* \in \mathcal{M}_{\mathbf{U}^*}$ is also a local minimizer.*

Proof. Assume this is not true for some $\hat{\mathbf{U}}^*$. Then there exists a sequence $(\hat{\mathbf{V}})_n$ converging to $\hat{\mathbf{U}}^* = \theta_{\mathbf{U}^*}(\mathsf{A})$ such that $j(\hat{\mathbf{V}}_n) > j(\hat{\mathbf{U}}^*)$ for all n. By the continuity of θ, the sequence $\mathbf{V}_n = \theta_{\hat{\mathbf{V}}_n}(\mathsf{A}^{-1})$ converges to $\mathbf{U}^*$ and satisfies $j(\mathbf{V}_n) = j(\hat{\mathbf{V}}_n) > j(\hat{\mathbf{U}}^*) = j(\mathbf{U}^*)$. This contradicts the local optimality of $\mathbf{U}^*$. ∎

In light of this observation, we will call $\mathcal{M}_{\mathbf{U}^*}$ a *local solution orbit* of (2.3). The alternating optimization approach to find a representative of $\mathcal{M}_{\mathbf{U}^*}$ consists of iterating the cycle

$$
\begin{aligned}
\mathbf{U}_1^{(n+1)} &= \operatorname*{argmin}_{\mathbf{V}_1} j(\mathbf{V}_1, \mathbf{U}_2^{(n)}, \dots, \mathbf{U}_d^{(n)}),\\
\mathbf{U}_2^{(n+1)} &= \operatorname*{argmin}_{\mathbf{V}_2} j(\mathbf{U}_1^{(n+1)}, \mathbf{V}_2, \mathbf{U}_3^{(n)} \dots, \dots, \mathbf{U}_d^{(n)}),\\
&\vdots\\
\mathbf{U}_d^{(n+1)} &= \operatorname*{argmin}_{\mathbf{V}_d} j(\mathbf{U}_1^{(n+1)}, \dots, \mathbf{U}_{d-1}^{(n+1)}, \mathbf{V}_d)
\end{aligned}
\tag{2.4}
$$

(where $\mathbf{U}_i, \mathbf{V}_i \in \mathbb{R}^{r_{i-1}\times n_i \times r_i}$). In the case of best approximation (in the Frobenius norm) this method is called *alternating least squares* (ALS), since every microstep in (2.4) is a linear least squares problem then. More generally, this type of method (more precisely its gradient version) is referred to as nonlinear block Gauss–Seidel, SOR, or relaxation method [38]. In fact, it locally equals the linear Gauss–Seidel iteration applied to the Hessian $j''(\mathbf{U}^*)$ up to second order terms [3, 31, 32].

For computational and also, as we will see later, for analytical reasons, it can be useful to rescale the iterates during the process. Given sequences of scaling operators $R_1^{(n)}, R_2^{(n)}, \dots, R_d^{(n)}$ of the form

$$R_i^{(n)}(\mathbf{U}) = \theta_{\mathbf{U}}(\mathsf{A}_i^{(n)}(\mathbf{U})),$$

the algorithm we now analyze is displayed as Algorithm 1.

To study Algorithm 1 in a precise mathematical framework, we shall first convince ourselves that every ALS microstep of (2.4) has a unique solution if the current microiterate is in $\mathcal{U}$, that is, satisfies the full rank condition (A.8).

PROPOSITION 2.2. *Let $\mathbf{U} \in \mathcal{U}$. Then for $i = 1, 2, \dots, d$ the linear maps*

$$P_{\mathbf{U},i}\colon \mathbb{R}^{r_{i-1}\times n_i \times r_i} \to \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}\colon \mathbf{V}_i \mapsto \tau(\mathbf{U}_1, \dots, \mathbf{U}_{i-1}, \mathbf{V}_i, \mathbf{U}_{i+1}, \dots, \mathbf{U}_d)$$

ALGORITHM 1. TT-ALS with rescaling.

Require: $\mathbf{U}^{(0)}$
  for $n = 0, 1, 2, \dots$ **do**
    for $i = 1, 2, \dots, d$ **do**
      1. Perform one ALS microstep:
$$\widetilde{\mathbf{U}}_i^{(n+1)} = \operatorname*{argmin}_{\mathbf{V}_i} j(\mathbf{U}_1^{(n+1)}, \dots, \mathbf{U}_{i-1}^{(n+1)}, \mathbf{V}_i, \mathbf{U}_{i+1}^{(n)}, \dots, \mathbf{U}_d^{(n)}),$$
$$S_i(\mathbf{U}^{(n)}) = (\mathbf{U}_1^{(n+1)}, \dots, \mathbf{U}_{i-1}^{(n+1)} \widetilde{\mathbf{U}}_i^{(n+1)}, \mathbf{U}_{i+1}^{(n)}, \dots, \mathbf{U}_d^{(n)}).$$
      2. Rescale TT representation:
$$\mathbf{U}^{(n+1)} = R_i^{(n)}(S_i(\mathbf{U}^{(n)})).$$
    end for
  end for

are injective. The operators

$$S_i \colon \mathcal{U} \to \overline{\mathcal{U}} \colon \mathbf{U} \mapsto \left(\mathbf{U}_1, \dots, \mathbf{U}_{i-1}, \operatorname*{argmin}_{\mathbf{V}_i} J(P_{\mathbf{U},i}(\mathbf{V}_i)), \mathbf{U}_{i+1}, \dots, \mathbf{U}_d \right)$$

are well-defined and continuously differentiable on $\mathcal{U}$.

Proof. By (A.3) and (A.6), the ith canonical unfolding of $\mathbf{X} = P_{\mathbf{U},i}(\mathbf{V}_i)$ reads

$$\mathbf{X}^{(i)} = (I_{n_i} \otimes \mathbf{U}^{\leq i-1}) \mathbf{V}_i^L \mathbf{U}^{\geq i+1}.$$

Since the matrices $(I_{n_i} \otimes \mathbf{U}^{\leq i-1})$ and $\mathbf{U}^{\geq i+1}$ are of full column and row rank, respectively, and since unfoldings are isomorphisms, the injectivity of $P_{\mathbf{U},i}$ follows. Consequently, the map $\mathbf{V}_i \mapsto J(P_{\mathbf{U},i}(\mathbf{V}_i))$ is strictly convex and coercive (recall that we assumed that for J), and hence possesses a unique global minimizer which depends smoothly on $\mathbf{U}$ (since J is C^2). In other words, the S_i are well-defined and continuously differentiable. ∎

Assuming for the moment that the iteration process does not leave $\mathcal{U}$ at any stage (we prove this in Lemma 2.4), we may write Algorithm 1 as

$$\mathbf{U}^{(n+1)} = (R_d^{(n)} \circ S_d \circ R_{d-1}^{(n)} \circ S_{d-1} \circ \cdots \circ R_1^{(n)} \circ S_1)(\mathbf{U}^{(n)}). \tag{2.5}$$

A crucial observation is that the true object of interest, the sequence of the TT tensors $\tau(\mathbf{U}^{(n)})$, does not depend on the choice of the scaling operations $R_i^{(n)}$. For illustration, let us recall the rank-one case. If a^* is a unique minimizer of $a \mapsto J(a \otimes b \otimes c)$ for fixed b, c, then $\beta^{-1}\gamma^{-1}a^*$ is the unique minimizer of $a \mapsto J(a \otimes \beta b \otimes \gamma c)$. Hence one ends up with equivalent representations $a^* \otimes b \otimes c$ and $\beta^{-1}\gamma^{-1}a^* \otimes \beta b \otimes \gamma c$. Interestingly, this generalizes to TT tensors.

PROPOSITION 2.3. *Let* $\mathbf{U} \in \mathcal{U}$ *and assume* $S_i(\mathbf{U}) \in \mathcal{U}$. *Then equivalent TT representations* $\hat{\mathbf{U}} \in \mathcal{M}_{\mathbf{U}}$ *are mapped to equivalent representations* $S_i(\hat{\mathbf{U}}) \in \mathcal{M}_{S_i(\mathbf{U})}$, *that is,*

$$S_i(\mathcal{M}_{\mathbf{U}}) \subseteq \mathcal{M}_{S_i(\mathbf{U})}.$$

Proof. The main argument is that the linear operators $P_{\mathbf{U},i}$ and $P_{\hat{\mathbf{U}},i}$ have the same range [21, Lemma 3.2], and that J, by virtue of its strict convexity, has a unique minimizer $\mathbf{X}$ on that space. Hence, denoting by $\mathbf{V}_i$ and $\hat{\mathbf{V}}_i$ the minimizers of $J \circ P_{\mathbf{U},i}$ and $J \circ P_{\hat{\mathbf{U}},i}$, respectively, it holds that

$$\tau(S_i(\mathbf{U})) = P_{\mathbf{U},i}(\mathbf{V}_i) = \mathbf{X} = P_{\hat{\mathbf{U}},i}(\hat{\mathbf{V}}_i) = \tau(S_i(\hat{\mathbf{U}})).$$

The assertion follows from Proposition A.4 in the form of (A.11). ∎

The preceding proposition identifies Algorithm 1 as an algorithm on the manifold $\mathcal{T}_{\mathbf{r}}$ of tensors of fixed TT rank. The value for the convergence analysis lies in the possibility to restrict the attention to particular choices of scaling strategies which are easy to investigate. Using this idea, we first can prove that $\mathbf{U}^{(n+1)}$ in (2.5) is uniquely defined.

LEMMA 2.4. *Let $\mathbf{U}^* \in \mathcal{U}$ be a local solution of minimization problem* (2.3)*. Then $\mathcal{M}_{\mathbf{U}^*}$ possesses an open neighborhood $\mathcal{O} \subseteq \mathcal{U}$ in which the composed operator*

$$R_d^{(n)} \circ S_d \circ R_{d-1}^{(n)} \circ S_{d-1} \circ \cdots \circ R_1^{(n)} \circ S_1 \colon \mathcal{O} \to \mathcal{U} \tag{2.6}$$

is well-defined for all sequences of scaling operators $(R_i^{(n)})$. Hence, the result of one full step of Algorithm 1 is uniquely defined if the current iterate $\mathbf{U}^{(n)}$ is in $\mathcal{O}$. The next iterate is given by (2.5)*.*

Furthermore, every $\hat{\mathbf{U}}^* \in \mathcal{M}_{\mathbf{U}^*}$ is a fixed point of every S_i. Consequently, it holds that

$$(R_d^{(n)} \circ S_d \circ R_{d-1}^{(n)} \circ S_{d-1} \circ \cdots \circ R_1^{(n)} \circ S_1)(\hat{\mathbf{U}}^*) = (R_d^{(n)} \circ R_{d-1}^{(n)} \circ \cdots \circ R_1^{(n)})(\hat{\mathbf{U}}^*).$$

Proof. We first show the second part. As we have seen in the proof of Proposition 2.2, the $S_i(\mathbf{U}^*)$ are constructed by replacing $\mathbf{U}_i^*$ with the unique global minimum of the strictly convex functional $\mathbf{V}_i \mapsto J \circ P_{\mathbf{U}^*,i}$. Since by assumption $\mathbf{U}_i^*$ has to be a local minimum of this function, it is already the global one. Due to Proposition 2.1, this argument works for all $\hat{\mathbf{U}}^* \in \mathcal{M}_{\mathbf{U}^*}$.

Set $\mathcal{O}_d = \mathcal{U}$. Since $\mathcal{U}$ is open and $\mathbf{U}^*$ is a fixed point of the continuous map S_{d-1}, there exists an open neighborhood $\mathcal{O}_{d-1}$ of $\mathbf{U}^*$ such that $S_{d-1}(\mathcal{O}_{d-1}) \subseteq \mathcal{O}_d$. Hence $S_d \circ S_{d-1}$ is defined on $\mathcal{O}_{d-1}$. By induction, we find a neighborhood $\mathcal{O}_1$ of $\mathbf{U}^*$ such that

$$S_d \circ S_{d-1} \circ \cdots \circ S_1$$

is defined on $\mathcal{O}_1$ and even, by Proposition 2.3, on

$$\mathcal{O} := \{\theta_{\mathbf{U}}(\mathsf{A}) : \mathbf{U} \in \mathcal{O}_1,\, \mathsf{A} \in \mathcal{G}\},$$

which is an open neighborhood of $\mathcal{M}_{\mathbf{U}^*}$. Proposition 2.3 also proves that operator (2.6) is well-defined on $\mathcal{O}$ for any choice of scaling operators $R_i^{(n)}$. ∎

2.2. Main assumption and convergence results. Formally, Algorithm 1 is an algorithm in the parameter space $\overline{\mathcal{U}}$ and produces a sequence $(\mathbf{U}^{(n)}) \subseteq \mathcal{U}$ if the starting point is close enough to a local solution $\mathbf{U}^* \in \mathcal{U}$ of the minimization problem (2.3). The standard approach to prove local convergence $\mathbf{U}^{(n)} \to \mathbf{U}^*$ would be a contraction argument. But since we made no assumptions on the scaling operators $R_i^{(n)}$ so far, no point of the solution orbit $\mathcal{M}_{\mathbf{U}^*}$ has to be a fixed point of the iteration at all. On the other hand, we initially started with the minimization problem (2.1) on the set $\mathcal{T}_{\leq \mathbf{r}}$. In practice, we are thus only interested in the convergence $\tau(\mathbf{U}^{(n)}) \to \tau(\mathbf{U}^*)$. According to Proposition 2.3 the latter sequence is independent of the scaling operators $R_i^{(n)}$. Our trick is now to use a scaling strategy that enables us to use fixed-point arguments in the parameter space.

As a particular instance of (2.5), we investigate the iteration

$$\mathbf{U}^{(n+1)} = (R_{\mathbf{U}^*} \circ S)(\mathbf{U}^{(n)}), \tag{2.7}$$

where $R_{\mathbf{U}^*}$ is a local normalization operator as described in Proposition A.6 and

$$S := S_d \circ S_{d-1} \circ \cdots \circ S_1$$

is the "pure" ALS operator. By Lemma 2.4, $R_{\mathbf{U}^*} \circ S$ is well-defined and smooth in a neighborhood of $\mathbf{U}^*$. Note that every point in $\mathcal{M}_{\mathbf{U}^*}$ is a fixed point of S, but only $\mathbf{U}^*$ is a fixed point of $R_{\mathbf{U}^*}$, at least locally. Hence $\mathbf{U}^*$ is the only fixed point of $R_{\mathbf{U}^*} \circ S$ in $\mathcal{M}_{\mathbf{U}^*}$ in a neighborhood of itself.

If j would be a quadratic form, then the alternating optimization given by operator S is just the standard block Gauss–Seidel iteration applied to the Hessian matrix $j''(\mathbf{U}^*)$ of j at the minimizer $\mathbf{U}^*$. If j is arbitrary, then this is at least asymptotically true (Lemma 2.6 below). Since j is constant on the orbit $\mathcal{M}_{\mathbf{U}^*}$, the null space of $j''(\mathbf{U}^*)$ at least contains the $(\sum_{i=1}^{d-1} r_i^2)$-dimensional tangent space $T\mathcal{M}_{\mathbf{U}^*}$ (see (A.12)). Thus, as known from theory, elements in $T\mathcal{M}_{\mathbf{U}^*}$ will not be damped by the Gauss–Seidel method. As we will see later, this drawback is compensated by the local normalization operator $R_{\mathbf{U}^*}$. What we need to assume for our convergence proof is that $j''(\mathbf{U}^*)$ is positive definite in all other directions. Such an assumption is natural and in line with usual results concerning the convergence of the nonlinear SOR, if one takes the scaling indeterminacy in our setting into account.

MAIN ASSUMPTION. *At the local minimizer* $\mathbf{U}^*$*, the Hessian* $j''(\mathbf{U}^*)$ *is of full possible rank,*

$$\text{(MA)} \qquad \operatorname{rank} j''(\mathbf{U}^*) = \sum_{i=1}^{d} r_{i-1} n_i r_i - \sum_{i=1}^{d-1} r_i^2, \quad \textit{that is,} \quad \ker j''(\mathbf{U}^*) = T\mathcal{M}_{\mathbf{U}^*}.$$

As one can think about, this assumption implies $\mathbf{U}^* \in \mathcal{U}$ as long as $j'(\mathbf{U}^*) = 0$, which we, however, state separately in the following theorems for clarity. We should make clear that the particular choice of $\mathbf{U}^*$ has no qualitative influence. In accordance to our viewpoint of the ALS algorithm as an "iteration of orbits," assumption (MA) is, in fact, an assumption on the whole solution orbit $\mathcal{M}_{\mathbf{U}^*}$.

PROPOSITION 2.5. *Let* $\mathcal{M}_{\mathbf{U}^*} \subseteq \mathcal{U}$ *be a local solution orbit. If* (MA) *holds for* $\mathbf{U}^* \in \mathcal{M}_{\mathbf{U}^*}$*, then it holds for all* $\hat{\mathbf{U}}^* \in \mathcal{M}_{\mathbf{U}^*}$.

Proof. Let $\hat{\mathbf{U}}^* = \theta_{\mathsf{A}^{-1}}(\mathbf{U}^*)$ for some $\mathsf{A} \in \mathcal{G}$.[1] Then for all $\mathbf{H} \in \overline{\mathcal{U}}$ we have

$$j(\hat{\mathbf{U}}^* + \mathbf{H}) = j(\theta_{\mathsf{A}^{-1}}(\mathbf{U}^* + \theta_{\mathsf{A}}(\mathbf{H})) = j(\mathbf{U}^* + \theta_{\mathsf{A}}(\mathbf{H})),$$

where we used that j is constant on orbits. Taking the linearity of θ_{A} into account, the above relation implies

$$j''(\hat{\mathbf{U}}^*)[\mathbf{H}, \mathbf{H}] = j''(\mathbf{U}^*)[\theta_{\mathsf{A}}(\mathbf{H}), \theta_{\mathsf{A}}(\mathbf{H})].$$

Now note that θ_{A} is an isomorphism from $\overline{\mathcal{U}}$ onto itself (with inverse $\theta_{\mathsf{A}^{-1}}$). Hence $j''(\hat{\mathbf{U}}^*)$ and $j''(\mathbf{U}^*)$ are of same rank. ∎

We now present the convergence analysis of Algorithm 1 under assumption (MA). The validity of this assumption will be discussed in the next section.

LEMMA 2.6. *Let* $\mathbf{U}^* \in \mathcal{U}$ *be a local minimum of* (2.3) *for which main assumption* (MA) *holds. Partition* $j''(\mathbf{U}^*) = L + D + U$ *according to the block variables* $\mathbf{U}_i$

[1] In this proof we use the shorthand θ_{A} for the map $\mathbf{U} \mapsto \theta_{\mathbf{U}}(\mathsf{A})$.

into lower block triangular, block diagonal, and upper block triangular matrices L, D, and U, respectively. Then D is positive definite and it holds that

$$S'(\mathbf{U}^*) = -(D+L)^{-1}U.$$

Proof. Clearly, $j''(\mathbf{U}^*)$ is positive semidefinite. Using the full rank properties (A.8) of $\mathbf{U}^*$, it is easily verified from (A.12) that vectors of the form $\mathbf{H} = (0,\dots,0,\mathbf{H}_i,0,\dots,0)$ do not belong to $T\mathcal{M}_{\mathbf{U}^*}$ unless $\mathbf{H}_i = 0$. It follows from (MA) that D is positive definite, and hence $D+L$ invertible. A nice calculation of $S'(\mathbf{U}^*)$ which leads to the asserted formula can be found in [3, Lemma 2]. The classical references are [32, sect. 4] or [31, sect. 10.3.4-5]. ∎

In the following, we denote by

$$|\mathbf{H}|_E = (j''(\mathbf{U}^*)[\mathbf{H},\mathbf{H}])^{1/2}$$

the energy seminorm of $j''(\mathbf{U}^*)$. The corresponding operator seminorm is $|T|_E = \sup_{|\mathbf{H}|_E} |T\mathbf{H}|_E$. Further, let $\|\cdot\|$ denote a norm on $\overline{\mathcal{U}}$. If main assumption (MA) holds, then

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

also defines a norm, for if $|\mathbf{H}|_E = 0$, then $\mathbf{H} \in \ker j''(\mathbf{U}^*) = T\mathcal{M}_{\mathbf{U}^*}$, and since $R_{\mathbf{U}^*}$ is constant on a neighborhood of $\mathbf{U}^*$ in $T\mathcal{M}_{\mathbf{U}^*}$, $\|(I - R'_{\mathbf{U}^*}(\mathbf{U}^*))\mathbf{H}\| = \|\mathbf{H}\| = 0$ implies $\mathbf{H} = 0$.

THEOREM 2.7. *Let $\mathbf{U}^* \in \mathcal{U}$ be a local minimum of* (2.3) *for which main assumption* (MA) *holds. Then the iteration* (2.7) *is locally q-linearly convergent to $\mathbf{U}^*$ in the norm $\|\cdot\|_*$ at an asymptotic rate of at least $q = |S'(\mathbf{U}^*)|_E < 1$. That is, for every $\epsilon > 0$ there exists a neighborhood $\mathcal{O}^*_\epsilon \subseteq \mathcal{U}$ of $\mathbf{U}^*$ such that*

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

for all n, if $\mathbf{U}^{(0)}$ is in $\mathcal{O}^*_\epsilon$.

Proof. Since $R^2_{\mathbf{U}^*} = R_{\mathbf{U}^*}$ in a neighborhood of $\mathbf{U}^*$, it holds that

$$(I - R'_{\mathbf{U}^*}(\mathbf{U}^*))R'_{\mathbf{U}^*}(\mathbf{U}^*) = 0.$$

Further, for sufficiently small $\mathbf{H} \in \overline{\mathcal{U}}$, we have $j(\mathbf{U}^* + \mathbf{H}) = j(R_{\mathbf{U}^*}(\mathbf{U}^* + \mathbf{H}))$, which together with $j'(\mathbf{U}^*) = j'(R_{\mathbf{U}^*}(\mathbf{U}^*)) = 0$ shows

$$j''(\mathbf{U}^*)[\mathbf{H},\mathbf{H}] = j''(\mathbf{U}^*)[R'_{\mathbf{U}^*}(\mathbf{U}^*)\mathbf{H}, R'_{\mathbf{U}^*}(\mathbf{U}^*)\mathbf{H}],$$

that is,

$$|R'_{\mathbf{U}^*}(\mathbf{U}^*)|_E = 1.$$

Consequently, we can estimate

$$\|(R_{\mathbf{U}^*} \circ S)'(\mathbf{U}^*)\mathbf{H}\|_* = \|R'_{\mathbf{U}^*}(\mathbf{U}^*)S'(\mathbf{U}^*)\mathbf{H}\|_* \le |S'(\mathbf{U}^*)|_E|\mathbf{H}|_E \le |S'(\mathbf{U}^*)|_E\|\mathbf{H}\|_*.$$

By Lemma 2.6, $S'(\mathbf{U}^*)$ is the error iteration matrix of the linear block Gauss–Seidel method applied to $j''(\mathbf{U}^*)$ and the block diagonal is positive definite. It is known that under this condition, $S'(\mathbf{U}^*)$ is a contraction in the energy seminorm, that is, $|S'(\mathbf{U}^*)|_E < 1$; see [46, Hilfssatz 1], [24, eq. (9)], or [29, Theorem 3.2]. The theorem is thus a consequence of the contraction principle. ∎

Generic estimates for $q = |S'(\mathbf{U}^*)|_E$ are given in [49].

We can now give two equivalent convergence statements for arbitrary scalings.

COROLLARY 2.8. *Let $\mathcal{M}_{\mathbf{U}^*} \subseteq \mathcal{U}$ be a local solution orbit of* (2.3) *for which the main assumption* (MA) *holds.*[2] *Then the sequence $(\mathcal{M}_{\mathbf{U}^{(n)}})$ of orbits produced by Algorithm* 1 *is locally r-linearly convergent to $\mathcal{M}_{\mathbf{U}^*}$ at an asymptotic rate of at least $q = |S'(\mathbf{U}^*)|_E < 1$. That is, for every $\epsilon > 0$ there exists a neighborhood $\mathcal{O}_\epsilon \subseteq \mathcal{U}$ of $\mathcal{M}_{\mathbf{U}^*}$ such that*

$$\limsup_{n\to\infty} \sqrt[n]{\operatorname{dist}(\mathcal{M}_{\mathbf{U}^*}, \mathcal{M}_{\mathbf{U}^{(n)}})} \leq q + \epsilon$$

if $\mathbf{U}^{(0)}$ is in $\mathcal{O}_\epsilon$. (The distance is measured in any norm $\|\cdot\|$ on $\overline{\mathcal{U}}$.)

Proof. Since it is independent from the choice of norm, the assertion follows immediately from Theorem 2.7 and Proposition 2.3 by choosing a suitable neighborhood $\mathcal{O}_\epsilon^*$ of any $\mathbf{U}^* \in \mathcal{M}_{\mathbf{U}^*}$ and setting $\mathcal{O}_\epsilon := \{\theta_{\mathbf{U}}(\mathsf{A}) : \mathbf{U} \in \mathcal{O}_\epsilon^*, \mathsf{A} \in \mathcal{G}\}$, . ∎

COROLLARY 2.9. *Let $\mathbf{X}^* = \tau(\mathbf{U}^*)$ be a local solution of* (2.1) *with full TT rank $\mathbf{r}$, that is, $\mathbf{U}^* \in \mathcal{U}$. Assume main assumption* (MA) *holds for $\mathbf{X}^*$.*[2] *Then the sequence $(\mathbf{X}^{(n)}) = (\tau(\mathbf{U}^{(n)}))$ of TT tensors produced by Algorithm* 1 *is locally r-linearly convergent to $\mathbf{X}^*$ at an asymptotic rate of at least $q = |S'(\mathbf{U}^*)|_E < 1$. That is, for every $\epsilon > 0$ there exists a neighborhood $\mathcal{X}_\epsilon \subseteq \mathcal{T}_{\leq \mathbf{r}}$ of $\mathbf{X}^*$ such that*

$$\limsup_{n\to\infty} \sqrt[n]{\|\mathbf{X}^* - \mathbf{X}^{(n)}\|} \leq q + \epsilon$$

if $\mathbf{X}^{(0)} = \tau(\mathbf{U}^{(0)})$ is in $\mathcal{X}_\epsilon$ (with $\|\cdot\|$ an arbitrary norm on $\mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$).

Proof. This follows from Theorem 2.7 and Proposition 2.3 by choosing a suitable neighborhood $\mathcal{O}_\epsilon^*$ of any $\mathbf{U}^* \in \mathcal{M}_{\mathbf{U}^*}$ and a constant $C_\epsilon > 0$ such that

$$\|\tau(\mathbf{U}^*) - \tau(\mathbf{U})\| \leq C_\epsilon \|\mathbf{U}^* - \mathbf{U}\|_*$$

for all $\mathbf{U} \in \mathcal{O}_\epsilon^*$. Then, by Proposition A.3, $\mathcal{X}_\epsilon := \tau(\mathcal{O}_\epsilon^*)$ is a neighborhood of $\mathbf{X}^*$ in $\mathcal{T}_{\mathbf{r}}$ for which the assertion holds. Since $\mathcal{T}_{\mathbf{r}}$ is open in $\mathcal{T}_{\leq \mathbf{r}}$, $\mathcal{X}_\epsilon$ is also a neighborhood in $\mathcal{T}_{\leq \mathbf{r}}$. ∎

2.3. Convergence of ALS from [21]. The alternating linear scheme (also ALS) as introduced in [21] is a symmetric extension of Algorithm 1. The algorithm is described in Algorithm 2, where we used the notation ten for the inverse operations of the left and right unfolding in the first and second inner loop, respectively.

As one can see, the components $\mathbf{U}_i$ are first optimized from left to right and then backwards. After each microstep the representation is changed according to a QR decomposition of the unfolding. Shifting of R and R^T was noted for the convenience of the reader, but need not be implemented, since the affected component will be updated in the next microstep anyway. After the first inner loop, the representation is left orthonormal, and after a full loop right orthonormal (see Appendix A.3). The details of the algorithm are described in [21].

The convergence analysis of Algorithm 2 proceeds by the same lines as for Algorithm 1. It can be shown that it is well-defined in a neighborhood of a full rank solution orbit $\mathcal{M}_{\mathbf{U}^*} \subseteq \mathcal{U}$. More precisely, let $S_L = S$ denote the left to right ALS operator from above, and $S_R = S_d \circ S_{d-1} \circ \cdots \circ S_1$ the right to left operator. Then Algorithm 2 produces the same orbits as

$$R_{\mathbf{U}^*} \circ S_R \circ S_L$$

[2] This phrase makes sense by Proposition 2.5.

ALGORITHM 2. ALS

Require: $\mathbf{U}^{(0)}$
for $n = 0, 1, 2, \dots$ **do**
for $i = 1, 2, \dots, d-1$ **do**
1. Perform one ALS microstep:
$\widetilde{\mathbf{U}}_i^{(n+1)} = \operatorname{argmin}_{\mathbf{V}_i} j(\widetilde{\mathbf{U}}_1^{(n+1)}, \dots, \widetilde{\mathbf{U}}_{i-1}^{(n+1)}, \mathbf{V}_i, \mathbf{U}_{i+1}^{(n)}, \dots, \mathbf{U}_d^{(n)})$.
2. Make a (tall) QR decomposition:
$(\widetilde{\mathbf{U}}_i^{(n+1)})^L = QR$.
3. Keep Q, shift R to the right:
$\widetilde{\mathbf{U}}_i^{(n+1)} \leftarrow \operatorname{ten}(Q), \qquad \mathbf{U}_{i+1}^{(n)} \leftarrow R\mathbf{U}_{i+1}^{(n)}$.
end for
for $i = d, d-1, \dots, 2$ **do**
1. Perform one ALS microstep:
$\mathbf{U}_i^{(n+1)} = \operatorname{argmin}_{\mathbf{V}_i} j(\widetilde{\mathbf{U}}_1^{(n+1)}, \dots, \widetilde{\mathbf{U}}_{i-1}^{(n+1)}, \mathbf{V}_i, \mathbf{U}_{i+1}^{(n+1)}, \dots, \mathbf{U}_d^{(n+1)})$.
2. Make a (tall) QR decomposition:
$((\mathbf{U}_i^{(n+1)})^R)^T = QR$.
3. Keep Q^T, shift R^T to the left:
$\mathbf{U}_i^{(n+1)} \leftarrow \operatorname{ten}(Q^T), \qquad \widetilde{\mathbf{U}}_{i-1}^{(n)} \leftarrow \widetilde{\mathbf{U}}_{i-1}^{(n)} R^T$.
end for
end for

(where formally $\mathbf{U}_d$ and $\mathbf{U}_1$ are updated twice in a row). The local behavior is governed by the spectral properties of the matrix $(S_R \circ S_L)'(\mathbf{U}^*) = S_R'(\mathbf{U}^*)S_L'(\mathbf{U}^*)$, which is the error iteration matrix of the symmetric linear block Gauss–Seidel iteration. In the same way as $S_L'(\mathbf{U}^*)$, matrix $S_R'(\mathbf{U}^*)$ is a contraction in the energy seminorm [29, Theorem 3.2], and hence

$$|(S_R \circ S_L)'(\mathbf{U}^*)|_E \le |S_R'(\mathbf{U}^*)|_E \cdot |S_L'(\mathbf{U}^*)|_E < 1.$$

We conclude with the following theorem.

THEOREM 2.10. *The convergence results of Theorem* 2.7 *and Corollaries* 2.8, 2.9 *hold for the ALS Algorithm* 2 *from* [21]*, with the convergence rate replaced by* $q = |(S_R \circ S_L)'(\mathbf{U}^*)|_E$.

2.4. Decomposition and approximation of tensors with known rank. Ideally, we would like to find a minimizer $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ of J. Assume we have chosen for $\mathbf{r}$ the correct TT rank of $\mathbf{Y}$ and that $\mathbf{U}^*$ is the solution of (2.3) with $\tau(\mathbf{U}^*) = \mathbf{Y}$. In this case it holds that $J'(\mathbf{Y}) = 0$ so that (cf. (3.3))

$$j''(\mathbf{U}^*)[\mathbf{H}, \mathbf{H}] = J''(\mathbf{Y})[\tau'(\mathbf{U}^*)\mathbf{H}, \tau'(\mathbf{U}^*)\mathbf{H}].$$

Using Proposition A.3 we can conclude the following.

THEOREM 2.11. *Assume the* $\mathbf{Y} = \tau(\mathbf{U}^*)$ *has TT rank* $\mathbf{r}$ *and that* $J''(\mathbf{Y})$ *is positive definite. Then the main assumption* (MA) *and the statements of Theorems* 2.7 *and* 2.10 *hold.*

This is a very pleasant result highlighting the importance of a correct rank estimation, which, however, is an open problem.

Please note that the assumption on $J''(\mathbf{Y})$ particularly holds for the best least squares approximation problem where $J'' = I$. However, in this case we can even show

a little more. In practice it is observed that, if the rank of the approximand $\mathbf{Y}$ matches that of the used TT manifold, the ALS from Algorithms 1 and 2 usually returns a TT decomposition of $\mathbf{Y}$ after one run over all components; see [21] for numerics. As a generalization of [21, Lemma 4.2], we prove that this holds independently of the scaling and of orthogonality of the components.

PROPOSITION 2.12. *Assume* $\mathbf{Y} = \tau(\mathbf{V}) \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ *has TT rank* $\mathbf{r}$. *Then Algorithm* 1 *applied to*

$$j \colon \overline{\mathcal{U}} \to \mathbb{R} \colon \mathbf{U} \mapsto \frac{1}{2}\|\mathbf{Y} - \tau(\mathbf{U})\|_F^2$$

finishes with a TT decomposition $\mathbf{Y} = \tau(\mathbf{U}^{(1)})$ *after one loop over* $i = 1, 2, \ldots, d$ *if the starting point* $\mathbf{U}^{(0)}$ *is in*

$$\mathcal{V} = \{\mathbf{U} \in \mathcal{U} \mid \operatorname{rank}(\mathbf{V}_{i+1}^R (\mathbf{U}_{i+1}^R)^T) = r_i \text{ for } i = 1, 2, \ldots, d-1\}.$$

This set is open and dense in $\overline{\mathcal{U}}$. *The complement* $\overline{\mathcal{U}} \setminus \mathcal{V}$ *has measure zero.*

Proof. Let $\mathbf{Y} = \tau(\mathbf{V})$ be a left orthonormal $\mathrm{TT}_{\leq \mathbf{r}}$ decomposition (see (A.13)). We set $(\mathbf{U}^{(1)})^{\leq 0} = \mathbf{V}^{\leq 0} = 1$ and assume $(\mathbf{U}^{(1)})^{\leq i-1} = \mathbf{V}^{\leq i-1}$ for some $1 \leq i \leq d$. Using (A.3) and (A.6), it holds for the ith microstep that

$$\begin{aligned}(\mathbf{U}_i^{(1)})^L &= \operatorname*{argmin}_{Z \in \mathbb{R}^{r_{i-1} n_i \times r_i}} \|(I_{n_i} \otimes \mathbf{V}^{\leq i-1})[\mathbf{V}_i^L \mathbf{V}^{\geq i+1} - Z \cdot (\mathbf{U}^{(0)})^{\geq i+1}]\|_F \\ &= \operatorname*{argmin}_{Z \in \mathbb{R}^{r_{i-1} n_i \times r_i}} \|\mathbf{V}_i^L \mathbf{V}^{\geq i+1} - Z \cdot (\mathbf{U}^{(0)})^{\geq i+1}\|_F,\end{aligned}$$

where we used that $(I_{n_i} \otimes \mathbf{V}^{\leq i-1})$ has orthonormal columns (see (A.6)). Hence,

$$(\mathbf{U}_i^{(1)})^L = \mathbf{V}_i^L A$$

with[3] $A = \mathbf{V}^{\geq i+1}((\mathbf{U}^{(0)})^{\geq i+1})^\dagger$. It is easy to deduce from (A.7), that, by the choice of $\mathbf{U}^{(0)}$, matrix $\mathbf{V}^{\geq i+1}((\mathbf{U}^{(0)})^{\geq i+1})^T$ and with that matrix A are invertible. In consideration of Proposition 2.3 we can assume $A = I_{r_i}$ (for $i = d$ this necessarily holds) and proceed by induction.

For completeness we prove the assertions on $\mathcal{V}$. While it is obvious that $\mathcal{V}$ is open, we could not find a reference for $\overline{\mathcal{U}} \setminus \mathcal{V}$ having measure zero. We, therefore, deduce it from the following lemma by choosing, for each i, $E = \ker \mathbf{V}_{i+1}^R$ and $s = r = r_i$ therein. ∎

LEMMA 2.13. *Let* $1 \leq s \leq r \leq n$, *and let* E *be an* $(n-r)$*-dimensional subspace of* $\mathbb{R}^n$. *Consider the set* $\Omega_s = \{A \in \mathbb{R}^{n \times s} \mid \operatorname{rank} A = s, \operatorname{ran} A \cap E = \{0\}\}$. *Then the complement* Ω_s^c *has (Borel) measure zero in* $\mathbb{R}^{n \times s}$.

Proof. We proceed by induction over s. For $s = 1$ the assertion is clear. Assume it holds for some $s < r$. Then, $\Omega_s^c \times \mathbb{R}^n$, interpreted as subset of $\mathbb{R}^{n \times (s+1)}$, is a null set by Fubini's theorem. It hence suffices to show that $\Omega_{s+1}^c \cap (\Omega_s \times \mathbb{R}^n)$ has (Borel) measure zero in $\mathbb{R}^{n \times (s+1)}$. But this follows again from Fubini's theorem, since for each $A_s \in \Omega_s$ the cut

$$\{a \in \mathbb{R}^n \mid \begin{bmatrix} A_s & a \end{bmatrix} \in \Omega_{s+1}^c\} = \{a \in \mathbb{R}^n \mid a \in \operatorname{ran} A_s \oplus E\}$$

is a (Borel) null set in $\mathbb{R}^n$ (because $\dim(\operatorname{ran} A_s \oplus E) < n$). ∎

[3] By $M^\dagger = M^T(MM^T)^{-1}$ we denote the (right) pseudoinverse of a matrix M with full row rank.

2.5. Nonconvex functionals and approximate relaxation. The strict convexity of the functional J was needed in Proposition 2.2 to guarantee the unique solvability of the optimization steps (2.4). Hence the requirement of strict convexity can be relaxed to the somewhat imprecise assumption that every step of (2.4) possesses a unique solution. Even more generally, it is only necessary that the operators S_i can be adequately defined in a neighborhood of a solution orbit $\mathcal{M}_{\mathbf{U}^*}$, such that $\mathbf{V}_i = S_i(\mathbf{U})$ solves

$$j_i'(\mathbf{U}_1, \dots, \mathbf{U}_{i-1}, \mathbf{V}_i, \mathbf{U}_{i+1}, \dots, \mathbf{U}_d) = 0. \tag{2.8}$$

for $\mathbf{V}_i$, where j_i is the ith partial derivative of j. This is possible if the block diagonal of $j''(\mathbf{U}^*)$ is positive definite [31, sect. 10.3.5]. As we have seen in the proof of Lemma 2.6, this is already covered by assumption (MA).

The question is how to find a solution to (2.8). One of the simplest ideas is to apply one Newton step and proceed with the approximate solutions. This kind of procedure is called approximate nonlinear relaxation in [38] and Gauss–Seidel–Newton method in [31]. Alternatively, the method can be described as follows: at each microstep of (2.4), the functional

$$\mathbf{V}_i \mapsto j_i'(\mathbf{U}^{n,i})\mathbf{V}_i + \frac{1}{2} j_{ii}''(\mathbf{U}^{n,i})[\mathbf{V}_i, \mathbf{V}_i]$$

is minimized instead of j, where $\mathbf{U}^{n,i} = (\mathbf{U}_1^{(n+1)}, \dots, \mathbf{U}_{i-1}^{(n+1)}, \mathbf{U}_i^{(n)}, \dots, \mathbf{U}_d^{(n)})$ denotes the current microiterate and j_{ii}'' the ith diagonal block of the Hessian. Again, a corresponding iteration operator $\hat{S}$ can be defined in a neighborhood of $\mathcal{M}_{\mathbf{U}^*}$ if the block diagonal of $j''(\mathbf{U}^*)$ is positive definite (which follows from (MA)), and it holds that $\hat{S}'(\mathbf{U}^*)$ is the error iteration matrix of the linear block Gauss–Seidel method for $j''(\mathbf{U}^*)$; see the proof of [31, paragraph 10.3.3].

Of course, it should be even possible to maintain the local convergence property by replacing one exact Newton step by one step with a suitable approximate inverse. In the linear case this means to solve the restricted linear least squares problems only approximately, e.g., using only some steps of an iterative method.

By this general approach we obtain a local ALS-like method for arbitrary functions J. If we restrict ourselves to rescale the iterates only after one complete loop by a local normalization operator $R_{\mathbf{U}^*}$, that is, if we consider the iteration

$$\mathbf{U}^{(n+1)} = (R_{\mathbf{U}^*} \circ \hat{S})(\mathbf{U}^{(n)}),$$

then we can immediately conclude that the iterates will converge linearly to $\mathbf{U}^*$, provided that the starting point is close enough to $\mathbf{U}^*$ and assumption (MA) holds.

For the practical aspects concerning the choice of local normalization operators, we refer the reader to Appendix A.3.

3. Estimates for the Hessian and discussion of the main assumption. We return to the case that J is strictly convex. The main assumption (MA) entering the proof of convergence in Theorem 2.7 is that the Hessian $j''(\mathbf{U}^*)$ is of full possible rank $\dim \overline{\mathcal{U}} - \sum_{i=1}^{d-1} r_i^2$, meaning that $j''(\mathbf{U}^*)$ is positive definite on any complementary subspace of $T\mathcal{M}_{\mathbf{U}^*}$. As we have seen in the context of Theorem 2.11, this is the case if $\mathbf{Y} = \tau(\mathbf{U}^*)$ is in $\mathcal{T}_{\mathbf{r}}$ and solves $J'(\mathbf{Y}) = 0$. However, usually we do not know the correct rank $\mathbf{r}$ in advance. The problem in showing that (MA) holds when $J'(\tau(\mathbf{U}^*)) \neq 0$ is that the convexity of the functional J is not necessarily inherited by the functional

j defined on the parameter space $\overline{\mathcal{U}}$, even if the redundancy on $T\mathcal{M}_{\mathbf{U}^*}$ is factorized out. The properties of j rather depend both on the functional under consideration and on the properties of the manifold $\mathcal{T}_{\leq \mathbf{r}}$ used; therefore, condition (MA) is usually not trivial to verify. We will give two sufficient conditions for the TT-case in this section: one simple, more generic, and possibly void result for convex functionals, and one more elaborate, but somewhat pessimistic, result for the concrete example of least-squares approximation problems, where a gap condition on the singular values of the unfoldings of the tensor to be approximated guarantees that (MA) holds. The role of gaps between singular values has already been observed for the Tucker format in [8]. For the canonical format, see [42].

To verify (MA), we will in both cases use the following equivalent criterion, based on the idea of the introduction of gauge conditions for the parameter space already used in [20]. Recall the left unfoldings defined in (A.5).

CRITERION FOR (MA). *For* $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_d) \in \mathcal{U}$, *define*

$$W_{\mathbf{U}} := \{\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_d) \in \overline{\mathcal{U}} \mid (\mathbf{W}_i^L)^T \mathbf{U}_i^L = 0 \text{ for } i = 1, 2, \dots, d-1\} \tag{3.1}$$

as the so-called gauge space at $\mathbf{U}$. *This is a subspace of dimension* $\dim \overline{\mathcal{U}} - \sum_{i=1}^{d-1} r_i^2$ *of the component space* $\overline{\mathcal{U}}$. *Therefore, if*

$$j''(\mathbf{U}^*) \text{ is positive definite on } W_{\mathbf{U}^*}, \tag{3.2}$$

then (MA) *is fulfilled. In fact, it holds that* $\overline{\mathcal{U}} = T\mathcal{M}_{\mathbf{U}^*} \oplus W_{\mathbf{U}^*}$ *so that* (3.2) *and* (MA) *are equivalent.*

One could use any other space complementary to $T\mathcal{M}_{\mathbf{U}^*}$, but the above choice of $W_{\mathbf{U}^*}$ will turn out to be useful. Since $j = J \circ \tau$, we have to show that

$$j''(\mathbf{U}^*)[\mathbf{W}, \mathbf{W}] = J''(\tau(\mathbf{U}^*))[\tau'(\mathbf{U}^*)\mathbf{W}, \tau'(\mathbf{U}^*)\mathbf{W}] + J'(\tau(\mathbf{U}^*))(\tau''(\mathbf{U}^*)[\mathbf{W}, \mathbf{W}]) \tag{3.3}$$

is positive for all nonzero $\mathbf{W} \in W_{\mathbf{U}^*}$. Typically, for instance in the case of best approximation, the first term on the left side will be positive for all such $\mathbf{W} \neq 0$, which means nothing else that J'' is positive definite on the tangent space of $\mathcal{T}_{\mathbf{r}}$ at $\tau(\mathbf{U}^*)$ (Proposition A.3). However, this condition is not sufficient, one also has to control the second term.

3.1. A generic condition for positive definiteness. The theorem we state below suggests that at least for local minimizers $\mathbf{U}^* \in \mathcal{U}$ of j, which are sufficiently close to the unique minimizer $\mathbf{Y}$ of J on $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, assumption (MA) can be expected to hold. For this, it is necessary that the approximation manifold $\mathcal{T}_{\mathbf{r}}$ is good enough, that is, close enough to $\mathbf{Y}$. What this means exactly is difficult to say. The result is presented mainly to examine in its proof the important quantities one has to take care of. As before, we denote by $\|\cdot\|$ arbitrary norms on $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ and $\mathbb{R}^{r_{i-1} \times n_i \times r_i}$.

THEOREM 3.1. *Let* $\mathbf{Y}$ *be the unique minimum of the strict convex functional* J *on* $(\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}, \|\cdot\|)$. *Assume that* $J''(\mathbf{Y})$ *is positive definite. Then for every* $\mathbf{U} \in \mathcal{U}$, *there exists a* $\Delta > 0$ *(depending on* $\mathbf{U}$ *and* J*) such that if* $\mathbf{Y}$ *is close enough to* $\tau(\mathbf{U})$,

$$\|\mathbf{Y} - \tau(\mathbf{U})\| < \Delta,$$

then the following hold:

(i) *Assumption* (MA) *is satisfied for* $j''(\mathbf{U})$.

(ii) *If* $\mathbf{U}$ *solves* $j'(\mathbf{U}) = 0$*, there is an open neighborhood* $\mathcal{O}$ *of* $\tau(\mathbf{U})$ *such that* $\tau(\mathbf{U})$ *is the unique minimizer of* J *on* $\mathcal{O} \cap \mathcal{T}_{\leq \mathbf{r}}$.

Proof. Let $\|\cdot\|$ also denote the norm on $\overline{\mathcal{U}}$. First, by assumption, there is an $\alpha > 0$ with

$$\|J''(\mathbf{Y})[\mathbf{W}, \mathbf{W}]\| \geq \alpha \|\mathbf{W}\|^2.$$

Second, on some fixed ball of radius ϵ around $\mathbf{X}$, J' is Lipschitz continuous, so that

$$\|J'(\mathbf{X}) - J'(\tau(\mathbf{U}))\| \leq \beta \|\mathbf{X} - \tau(\mathbf{U})\|$$

for some $\beta > 0$ and all $\mathbf{X}$ in that ball. Further, it is not difficult to show that $\tau'(\mathbf{U})$ is injective on $W_{\mathbf{U}}$ (cf. Proposition A.3; this follows from Lemma 3.2 below). Hence there exists $\gamma > 0$ with

$$\|\tau'(\mathbf{U})\mathbf{W}\|^2 \geq \gamma \|\mathbf{W}\|^2$$

for all $\mathbf{W} \in W_{\mathbf{U}}$. Finally, we have

$$\|\tau''(\mathbf{U})[\mathbf{W}, \mathbf{W}]\| \leq \delta \|\mathbf{W}\|^2$$

for some $\delta > 0$. Now if

$$\|\mathbf{Y} - \tau(\mathbf{U})\| < \Delta = \min\left\{\frac{\alpha\gamma}{\beta\delta}, \epsilon\right\}, \tag{3.4}$$

then we can estimate (3.3) as follows:

$$\begin{aligned} j''(\mathbf{U})[\mathbf{W}, \mathbf{W}] &\geq \alpha \|\tau'(\mathbf{U})\mathbf{W}\|^2 - \|J'(\tau(\mathbf{U})) - J'(\mathbf{Y})\| \, \|\tau''(\mathbf{U})[\mathbf{W}, \mathbf{W}]\| \\ &\geq (\alpha\gamma - \beta\delta\Delta)\|\mathbf{W}\|^2 > 0 \end{aligned}$$

for all $\mathbf{W} \in W_{\mathbf{U}}$, where we used $J'(\mathbf{Y}) = 0$. By (3.2), this shows (i).

In particular, $\mathbf{U}$ is the unique solution of $j'(\mathbf{U}) = 0$ on some neighborhood of $\mathbf{U}$ in $\mathbf{U} + W_{\mathbf{U}}$. We again have to refer to Proposition A.3 which implies that τ is a diffeomorphism between such a neighborhood and a neighborhood of $\tau(\mathbf{U})$ in $\mathcal{T}_{\mathbf{r}}$. Hence $\tau(\mathbf{U})$ is the unique minimizer of J in a neighborhood of itself in $\mathcal{T}_{\mathbf{r}}$. Since this set is open in $\mathcal{T}_{\leq \mathbf{r}}$, (ii) follows. ∎

In this theorem, the critical distance Δ between $\tau(\mathbf{U})$ and $\mathbf{Y}$ depended on J, but also on the constants γ, δ bounding the derivatives of τ at $\mathbf{U}$, making it a potentially void statement. One has to ask whether Δ can be related to the minimizer $\mathbf{Y}$ only. As we will show in section 3.3, the answer is positive in the case of *best* least-square approximations. In any case, as in the previous proof, we need a lower bound for $\tau'(\mathbf{U})$ and an upper bound for $\tau''(\mathbf{U})$, at least in terms of $\tau(\mathbf{U})$.

3.2. Estimates for the derivatives of $\boldsymbol{\tau}$. We consider a general $\mathbf{U} \in \mathcal{U}$ and focus on the Frobenius norm $\|\cdot\|_F$. Note that $\|\mathbf{W}\|_F^2 = \sum_{i=1}^d \|\mathbf{W}_i\|_F^2$.

By Proposition 2.5, the condition (MA) transfers from any $\mathbf{U} \in \mathcal{U}$ to its orbit $\mathcal{M}_{\mathbf{U}}$. Therefore, we are free to choose a representation. Denoting by

$$\zeta = \min_{i=1,\dots,d-1} \sigma_{\min}(\tau(\mathbf{U})^{(i)}), \quad Z = \max_{i=1,\dots,d-1} \sigma_{\max}(\tau(\mathbf{U})^{(i)})$$

the smallest and largest singular value among all unfoldings of $\tau(\mathbf{U})$, respectively, we choose a weighted left orthogonal representation which satisfies

$$(\mathbf{U}^{\leq i})^T \mathbf{U}^{\leq i} = \zeta^{2i/d} I_{n_i}, \tag{3.5}$$

that is, the smallest singular value ζ is equally distributed among the left orthogonal factors $\mathbf{U}_i$. By our assumption that $\mathbf{U} \in \mathcal{U}$, ζ is positive. By (A.3) and (3.5), the singular values of $\mathbf{U}^{\geq i+1}$ are now contained in the interval $[\zeta^{(d-i)/d}, \zeta^{-i/d}Z]$.

By the multilinearity of τ, we have

$$\tau'(\mathbf{U})[\mathbf{W}] = \sum_{i=1}^{d} \tau(\mathbf{U}_1, \dots, \mathbf{W}_i, \dots, \mathbf{U}_d) \tag{3.6}$$

and

$$\tau''(\mathbf{U})[\mathbf{W}, \mathbf{W}] = 2\sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \tau(\mathbf{U}_1, \dots, \mathbf{W}_i, \dots, \mathbf{W}_j, \dots, \mathbf{U}_d). \tag{3.7}$$

We first consider the norms of the single terms by using isomorphy to suitable unfoldings (A.3) in combination with (A.6) and (A.7). It holds that

$$\|\tau(\mathbf{U}_1, \dots, \mathbf{W}_i, \dots, \mathbf{U}_d)\|_F = \|(I_{n_i} \otimes \mathbf{U}^{\leq i-1})\mathbf{W}_i^L \mathbf{U}^{\geq i+1}\|_F \geq \zeta^{(d-1)/d}\|\mathbf{W}_i\|_F. \tag{3.8}$$

To estimate the terms of (3.7) we need the matrices

$$\mathbf{U}^{\geq i+1}(\mathbf{W}_j) := [U_i(x_{i+1}) \dots W_j(x_j) \dots U_d(x_d)] \in \mathbb{R}^{r_i \times n_{i+1} \dots n_d} \tag{3.9}$$

for $i+1 \leq j \leq d$, which, using (3.5), satisfy

$$\begin{aligned}\|\mathbf{U}^{\geq i+1}(\mathbf{W}_j)\|_F &= \|\zeta^{-i/d}\mathbf{U}^{\leq i}\mathbf{U}^{\geq i+1}(\mathbf{W}_j)\|_F \\ &= \zeta^{(j-i-1)/d}\|\zeta^{-(j-1)/d}\mathbf{U}^{\leq j-1}\mathbf{U}^{\geq j}(\mathbf{W}_j)\|_F \\ &= \zeta^{(j-i-1)/d}\|\mathbf{U}^{\geq j}(\mathbf{W}_j)\|_F.\end{aligned}$$

Using this, it is easy to see that

$$\begin{aligned}\|\tau(\mathbf{U}_1, \dots, \mathbf{W}_i, \dots, \mathbf{W}_j, \dots, \mathbf{U}_d)\|_F &\leq \max(\zeta^{(d-2/d)}, \zeta^{-2/d}Z)\|\mathbf{W}_i\|_F\|\mathbf{W}_j\|_F \\ &= \zeta^{-2/d}Z\|\mathbf{W}_i\|_F\|\mathbf{W}_j\|_F.\end{aligned}$$

Now the advantages of the gauge space (3.1) come into play.

LEMMA 3.2. *Let $\mathbf{W} \in W_{\mathbf{U}}$. Then the terms in* (3.6) *are pairwise orthogonal in the Frobenius inner product. The same holds for two terms in* (3.7) *with different values of i.*

The proof is immediate from (3.5). By this lemma, (3.8) implies

$$\|\tau'(\mathbf{U})\mathbf{W}\|_F^2 \geq \gamma\|\mathbf{W}\|_F^2, \quad \gamma = \zeta^{2(d-1)/d}. \tag{3.10}$$

Next, again by Lemma 3.2, the Cauchy–Schwarz inequality, and (3.9),

$$\begin{aligned}\|\tau''(\mathbf{U})[\mathbf{W}, \mathbf{W}]\|^2 &\leq 4\zeta^{-4/d}Z^2 \sum_{i=1}^{d-1}(d-i)^2 \sum_{j=i+1}^{d} \|\mathbf{W}_i\|_F^2\|\mathbf{W}_j\|_F^2 \\ &= 2\zeta^{-4/d}Z^2 \sum_{i=1}^{d}\sum_{j=1}^{d} M(d)_{ij}\|\mathbf{W}_i\|_F^2\|\mathbf{W}_j\|_F^2 \\ &\leq 2\zeta^{-4/d}Z^2\|M(d)\|_2\|\mathbf{W}\|^4,\end{aligned}$$

where $\|M(d)\|_2$ is the spectral norm of the matrix with entries $(1-\delta_{ij})(d-\min(i,j))^2$. Since $M(d)$ is symmetric, its spectral norm is bounded by any induced matrix norm [22], in particular

$$\|M(d)\|_2 \leq \|M(d)\|_\infty = (d-1)^3.$$

We conclude with

$$\|\tau''(\mathbf{U})[\mathbf{W},\mathbf{W}]\|_F \leq \delta \|\mathbf{W}\|_F^2, \quad \delta = \sqrt{2}(d-1)^{3/2}\zeta^{-2/d}Z \tag{3.11}$$

for the weighted left orthogonal representations (3.5).

It is interesting to compare the results to estimates we would have obtained by using a pure left orthonormal representation (A.13). In that case, all of the spectral information would sit in the last component $\mathbf{U}_d$, so we lose it when varying this component. In particular, we would obtain $\gamma' = \min(1,\zeta^2)$, which is always less than $\zeta^{2(d-1)/d}$, and $\delta' = \max(1,Z)$, which appears better than δ in (1). However, in the following we will be interested in the ratio, and γ'/δ' would be difficult to estimate further, since it does not scale properly with the norm of $\tau(\mathbf{U})$.

3.3. Singular value gap conditions for least squares approximation. We now again consider

$$j(\mathbf{U}) = \frac{1}{2}\|\mathbf{Y} - \tau(\mathbf{U})\|_F^2 = \min, \tag{3.12}$$

where we have

$$j''(\mathbf{U})[\mathbf{W},\mathbf{W}] = \langle \tau'(\mathbf{U})\mathbf{W}, \tau'(\mathbf{U})\mathbf{W}\rangle_F + \langle \mathbf{Y} - \tau(\mathbf{U}), \tau''(\mathbf{U})[\mathbf{W},\mathbf{W}]\rangle_F. \tag{3.13}$$

Let $\tau(\mathbf{U}^*) \in \mathcal{T}_\mathbf{r}$ denote a *best* approximation of $\mathbf{Y}$ in $\mathcal{T}_{\leq \mathbf{r}}$. From (3.13) and Proposition A.3 it becomes perfectly clear that (MA) holds if $\tau(\mathbf{U}^*) = \mathbf{Y}$. On the other hand, it is clear that for any $\mathbf{Y}$ close enough to some $\widetilde{\mathbf{Y}} \in \mathcal{T}_\mathbf{r}$ all best approximations $\tau(\mathbf{U}^*)$ of $\mathbf{Y}$ in $\mathcal{T}_{\leq \mathbf{r}}$ will be sufficiently close to $\widetilde{\mathbf{Y}}$ by continuity of distance. Hence, since (3.13) is continuous in both $\mathbf{Y}$ and $\tau(\mathbf{U})$, it follows that $\widetilde{\mathbf{Y}}$ possesses a neighborhood such that for all $\mathbf{Y}$ from that neighborhood (MA) holds for all of its *best* approximations in $\mathcal{T}_{\leq \mathbf{r}}$. Unifying over all $\widetilde{\mathbf{Y}}$ we even obtain a whole neighborhood of $\mathcal{T}_\mathbf{r}$ with this property. This, by the way, necessarily means that the TT rank of a $\mathbf{Y}$ from this neighborhood is greater than or equal to $\mathbf{r}$, which we from now on assume.

We would like to quantify the effect using Theorem 3.1. Namely, for (3.12) we can choose $\alpha = \beta = 1$ and ϵ arbitrarily in the proof there. Consequently, by (3.4), (3.10), and (3.11), we need only achieve

$$\|\mathbf{Y} - \tau(\mathbf{U})\|_F < \Delta = \frac{\gamma}{\delta} = \frac{\zeta^2}{\sqrt{2}(d-1)^{3/2}Z} \tag{3.14}$$

for (MA) to hold, where ζ and Z are again the smallest and largest singular value among all unfoldings of $\tau(\mathbf{U})$. For brevity we set

$$C(d) = \frac{1}{\sqrt{2}(d-1)^{3/2}}. \tag{3.15}$$

Let $Y_{r_i}^{(i)}$ denote a best rank-r_i approximation of the ith unfolding $\mathbf{Y}^{(i)}$ of $\mathbf{Y}$. Further, let ι be an index for which ζ is the smallest singular value of $\tau(\mathbf{U})^{(\iota)}$. As an abbreviation, set

$$\varrho_\iota = \|\mathbf{Y}^{(\iota)} - Y_{r_\iota - 1}^{(\iota)}\|_F,$$

which is positive by assumption on $\mathbf{Y}$.

LEMMA 3.3. *Suppose that* $\mathbf{Y}$ *has TT rank larger than or equal to* $\mathbf{r}$*, and*

$$\|\mathbf{Y}-\tau(\mathbf{U})\|_F < \left(\varrho_\iota + \frac{Z}{2C(d)}\right) - \sqrt{\left(\varrho_\iota + \frac{Z}{2C(d)}\right)^2 - \varrho_\iota^2} \quad (< \varrho_\iota). \tag{3.16}$$

Then (3.14) *is satisfied.*

Proof. Using elementary calculus one finds that the condition is equivalent to

$$\|\mathbf{Y}-\tau(\mathbf{U})\|_F < \frac{C(d)}{Z}(\varrho_\iota - \|\mathbf{Y}-\tau(\mathbf{U})\|_F)^2, \quad \|\mathbf{Y}-\tau(\mathbf{U})\|_F < \varrho_\iota.$$

This is more restrictive than (3.14), since

$$\varrho_\iota = \inf_{\operatorname{rank} A = r_\iota - 1} \|\mathbf{Y}^{(\iota)} - A\|_F \le \|\mathbf{Y}-\tau(\mathbf{U})\|_F + \inf_{\operatorname{rank} A = r_\iota - 1} \|\tau(\mathbf{U})^{(\iota)} - A\|_F$$

and the latter infimum equals ζ. ∎

As a first step to get free from the dependence of $\tau(\mathbf{U})$, we are willing to sacrifice accuracy by replacing ϱ_ι by its lower bound

$$\varrho = \min_{i=1,\dots,d} \|\mathbf{Y}^{(i)} - Y^{(i)}_{r_i-1}\|_F \tag{3.17}$$

in this argumentation, which we from now on do (one expects $\varrho = \varrho_\iota$ in many cases where $\tau(\mathbf{U})$ is close to $\mathbf{Y}$). A first order expansion of the square root in (3.16) (with ϱ instead of ϱ_ι) leads to the even simpler sufficient condition[4]

$$\|\mathbf{Y}-\tau(\mathbf{U})\|_F \le \frac{\varrho^2}{2\left(\varrho + \frac{Z}{2C(d)}\right)} = \frac{C(d)\varrho^2}{2C(d)\varrho + Z} \quad (< \varrho). \tag{3.18}$$

Depending on ϱ, the condition becomes quite restrictive for growing d when $C(d)$ goes to zero. Better bounds for $\tau''(\mathbf{U})$ might help here a bit. At the level of the ith unfolding, the given bound means that the error obtained by using rank r_i is much smaller compared to the error one would obtain using a smaller rank. It is intuitively clear that there hence have to be considerable gaps between the r_ith and (r_i+1)th singular values of the unfoldings $\mathbf{Y}^{(i)}$.

From a different point of view, the above conditions allow us to construct for every $\tau(\mathbf{U}) \in \mathcal{T}_\mathbf{r}$ a neighborhood (depending on Z and hence $\tau(\mathbf{U})$) such that for all functions (3.12) with $\mathbf{Y}$ from that neighborhood assumption (MA) will hold for $\tau(\mathbf{U})$. Of course, what we would like to have instead is a criterion of how close a given $\mathbf{Y}$ has to be to the set of $\mathrm{TT}_{\le \mathbf{r}}$ tensors to guarantee (MA) for the best or quasi-best approximations. The main idea to derive such conditions is to link aforementioned singular value gap conditions for $\mathbf{Y}$ to the error of an high-order SVD (HOSVD) projection $\tau(\mathbf{U}) = \mathbf{Y}^{\mathrm{HOSVD}}$ of $\mathbf{Y}$. These kinds of projections are generalizations of the high-order singular value projection of De Lathauwer et al. for the Tucker format [5] to the TT format, and are obtained by some truncation scheme of the SVD expansions of corresponding unfoldings of $\mathbf{Y}$ after the r_ith position. A sequential version of this procedure is due to Oseledets and Tyrtyshnikov who call it TT-SVD algorithm [33], so to fix notation we take $\mathbf{Y}^{\mathrm{HOSVD}}$ to be the result of this algorithm. Without going

[4] By concavity of $f(x) = \sqrt{x}$, $\sqrt{x+y} = f(x+y) < f(x) + f'(x)y = \sqrt{x} - y/(2\sqrt{x})$ for $x > 0$.

into further details [36, 33, 16], we remark that $\mathbf{Y}^{\mathrm{HOSVD}}$ is a quasi-optimal $\mathrm{TT}_{\leq \mathbf{r}}$ approximation of $\mathbf{Y}$, can be written as the result of applying successively orthogonal projections to $\mathbf{Y}$, and satisfies

$$\|\mathbf{Y} - \mathbf{Y}^{\mathrm{HOSVD}}\|_F \leq \sqrt{\sum_{i=1}^{d-1} \|\mathbf{Y}^{(i)} - Y_{r_i}^{(i)}\|_F^2} = \sqrt{\sum_{i=1}^{d-1} \sum_{k_i > r_i} (\sigma_{k_i}^{(i)})^2}, \tag{3.19}$$

where by $\sigma_1^{(i)} \geq \sigma_2^{(i)} \geq \cdots$ we denote the singluar values of the ith unfolding $\mathbf{Y}^{(i)}$. The advantage of (3.19) for us is that the error only depends on $\mathbf{Y}$.

Now let

$$\Sigma = \max_{i=1,\ldots,d-1} \sigma_{\max}(\mathbf{Y}^{(i)}) \tag{3.20}$$

denote the largest singular value among all unfoldings of $\mathbf{Y}$. Then it holds for $\tau(\mathbf{U}) = \mathbf{Y}^{\mathrm{HOSVD}}$, since it is the result of succesive orthogonal projections, that $Z \leq \Sigma$. Therefore, we obtain from (3.18) and (3.19) our main result.

THEOREM 3.4. *Let* $\mathbf{Y}$ *be given with TT rank at least* $\mathbf{r}$ *and consider* j *as in* (3.12). *Assume that*

$$\sqrt{\sum_{i=1}^{d-1} \|\mathbf{Y}^{(i)} - Y_{r_i}^{(i)}\|_F^2} \leq \frac{C(d)\varrho^2}{2C(d)\varrho + \Sigma} \tag{3.21}$$

with $C(d)$ *from* (3.15), ϱ *from* (3.17), *and* Σ *from* (3.20). *Then assumption* (MA) *holds for all* $\tau(\mathbf{U})$ *which are closer to* $\mathbf{Y}$ *than* $\mathbf{Y}^{\mathrm{HOSVD}}$, *in particular for the best approximation* $\tau(\mathbf{U}^*)$ *in* $\mathcal{T}_{\leq \mathbf{r}}$.

As remarked earlier, the validity of (MA) already implies $\mathbf{U}^* \in \mathcal{U}$ (but maybe not for the other ones). The result is not sharp as we simplified the bound in Lemma 3.3 for convenience, from which we could obtain a tighter bound.

To see what Theorem 3.4 has to do with gap conditions, let $\mathbf{Y}$ have TT rank $\mathbf{s} \geq \mathbf{r}$. Further, let

$$\sigma = \min_{i=1,\ldots,d-1} \sigma_{r_i}^{(i)}, \quad \varsigma = \max_{i=1,\ldots,d-1} \sigma_{r_i+1}^{(i)}.$$

Then we can make the following rough, but sharp, estimates:

$$\sqrt{\sum_{i=1}^{d-1} \|\mathbf{Y}^{(i)} - Y_{r_i}^{(i)}\|_F^2} \leq \left(\sum_{i=1}^{d-1} s_i - r_i\right)^{1/2} \varsigma \quad \text{and} \quad \sigma \leq \varrho.$$

A stricter version of (3.21) hence is[5]

$$\varsigma \leq \left(\sum_{i=1}^{d-1} s_i - r_i\right)^{-1/2} \left(\frac{C(d)}{2C(d) + (\Sigma/\sigma)}\right) \sigma, \tag{3.22}$$

which we consider as a very pessimistic requirement due to the behavior of $C(d)$. Still, it is interesting to see the ratio Σ/σ entering in (3.22), suggesting that we obtain the

[5]The right-hand side of (3.21) is a monotonically increasing function w.r.t. $\varrho > 0$.

weakest gap requirement, if there are no large gaps in the singular values we would like to keep. However, this heuristic is only convincing as long as σ is (relatively) "large" and not Σ (relatively) "small." But this again would imply a very good proximity of $\mathbf{Y}$ by rank $\mathbf{r}$, since the norm square of $\mathbf{Y}$ is the sum of all squared singular values (per unfolding), which is fixed. Consequently, a gap between critical singular values is then automatically enforced.

We finish by highlighting the following special cases.

1. For reconstruction problems in which the rank of the tensor $\mathbf{Y}$ to be approximated is known to equal $\mathbf{r}$, we have $\|\mathbf{Y}^{(i)} - Y_{r_i}^{(i)}\|_F = 0$ so that (3.21) trivially holds. In that case, assumption (MA) follows directly from (3.13) (cf. Theorem 2.11). However, recall again that Algorithm 1 is usually finite and exact in that case (Proposition 2.12).
2. For the matrix case $d = 2$, already (3.18) implies $\|\mathbf{Y} - Y_r\| < \varrho$, which means

$$\sigma_{r+1} < \sigma_r,$$

and which, in turn, is equivalent to the uniqueness of the best rank-r approximation Y_r to $\mathbf{Y}$. It is not too difficult to see that uniqueness is necessary for (MA) to hold. Namely, if it is not satisfied, we can find a curve on the level set of best approximations which not coincides with $Y_r = (\mathbf{U}_1^*)^L A^{-1} A (\mathbf{U}_2^*)^R$.

For rank-one tensors, somewhat better statements have been obtained in [42] (cf. discussion at the end of section 3.2).

4. Conclusions and perspectives. We have shown local linear convergence of the ALS for the TT format, supplementing the linear convergence behavior of TT-ALS as observed in practice [21] by an according theoretical analysis. The proof bases on the convergence of the nonlinear Gauss–Seidel method, and shows that convergence does not depend on a specific component realization $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_d)$ of a tensor $\tau(\mathbf{U})$, but rather on the orbits $\mathcal{M}_{\mathbf{U}}$ of equivalent TT representations. As detailed in section 2.3, an important consequence is that the proof, therefore, also holds for the ALS-QR algorithm as proposed in [21]. There, a QR orthogonalization step is performed after each component optimization step each to keep the resulting equations for the components $\mathbf{U}_i$ well-conditioned, and we found this step to be an essential ingredient in view of the practical applicability of the ALS algorithm.

Our general idea of proof as pursued in this work extends [43] to the (classical and hierarchical) Tucker format and, more generally, to those tensor networks [10] for which the redundancy of the respective parametrization can be characterized explicitly as exemplified in this work for the TT format. Also, it should be investigated whether on the basis of our present proof convergence of the promising DMRG algorithm [47, 44, 39], used for eigenvalue computations in quantum physics and investigated lately under the acronym MALS for more general optimization tasks in [21], also could be verified.

The major task that remains in the TT case as well as in others is probably to verify the main assumption, i.e., the rank condition (MA) on the Hessian j'' of the composed functional $j = J \circ \tau$. For the TT case, we showed here that this assumption holds true, when the rank has been correctly estimated, or, at least in the case of best approximation, when $\mathbf{Y}$ is close enough to the manifold of chosen TT rank $\mathbf{r}$. The latter property is strongly related to gap conditions for the singular values of the different unfoldings of $\mathbf{Y}$ at the target ranks. Not surprisingly, finding generic conditions under which the assumption holds for general J appears to be a rather nontrivial task. One has to study the particular case.

Note that the condition (MA) also is on the very basis of proofs of convergence for many other algorithms that are or might be used in tensor optimization, examples including, for instance, the nonlinear (parallelizable) Jacobi method for the treatment of optimization tasks, or various variants of Newton's method applied to the gauge space characterized in [20] (see also [8] for an analogous approach for the Tucker format). As well, (MA) implies positive answers to important theoretical questions as existence of local best approximations on manifolds, local uniqueness of solutions of more global convex optimization tasks etc. Therefore, further necessary or sufficient conditions giving a characterization of cases where (MA) holds are strongly desirable.

Appendix.

A.1. Unfoldings and TT rank. If we treat a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ as a matrix

$$\mathbf{X}^{(i)} \in \mathbb{R}^{(n_1 \cdots n_i) \times (n_{i+1} \cdots n_d)},$$

then this is called the ith canonical unfolding of $\mathbf{X}$. Its matrix rank is called the ith separation rank of $\mathbf{X}$. To fix the ordering of multi-indices we choose the reverse lexicographical ordering[6] for $(n_1, \dots, n_i)$, and lexicographical ordering for $(n_{i+1}, \dots, n_d)$.

For $\mathbf{U} \in \overline{\mathcal{U}}$ and $i = 1, 2, \dots, d$ we further define the matrices

$$\mathbf{U}^{\leq i} = [U_1(x_1)U_2(x_2)\dots U_i(x_i)] \in \mathbb{R}^{n_1 n_2 \dots n_i \times r_i} \tag{A.1}$$

and

$$\mathbf{U}^{\geq i} = [U_i(x_i)U_{i+1}(x_{i+1})\dots U_d(x_d)] \in \mathbb{R}^{r_{i-1} \times n_i n_{i+1} \dots n_d}. \tag{A.2}$$

In $\mathbf{U}^{\leq i}$ the row vectors $U_1(x_1)U_2(x_2)\dots U_i(x_i)$ are stacked below each other in reverse lexicographical order. The columns $U_i(x_i)U_{i+1}(x_{i+1})\dots U_d(x_d)$ of $\mathbf{U}^{\geq i}$ are arranged in lexicographical order. According to (1.1), the ith unfolding of a $\mathrm{TT}_{\leq \mathbf{r}}$ tensor $\mathbf{X} = \tau(\mathbf{U})$ can be written as

$$\mathbf{X}^{(i)} = \mathbf{U}^{\leq i}\mathbf{U}^{\geq i+1}. \tag{A.3}$$

Hence, the ith separation rank of a $\mathrm{TT}_{\leq \mathbf{r}}$ tensor $\mathbf{X} = \tau(\mathbf{U})$ is at most r_i. On the other hand, if $\operatorname{rank} \mathbf{X}^{(i)} \leq r_i$ for $i = 1, 2, \dots, d-1$, it is always possible, using the successive TT-SVD algorithm of Oseledets [33], to find $\mathbf{U} \in \overline{\mathcal{U}}$ such that $\mathbf{X} = \tau(\mathbf{U})$ is a $\mathrm{TT}_{\leq \mathbf{r}}$ decomposition. This makes the following definition meaningful.

Definition A.1 (TT rank). *A tensor* $\mathbf{X}$ *has* TT rank $\mathbf{r} = (r_1, r_2, \dots, r_{d-1})$ *if* $\operatorname{rank} \mathbf{X}^{(i)} = r_i$ *for* $i = 1, 2, \dots, d-1$. *The values* r_i *are called* TT ranks *of* $\mathbf{X}$.

Note that we have the trivial bounds

$$r_i \leq \min\left(\prod_{j=1}^{i} n_j, \prod_{j=i+1}^{d} n_j\right) \tag{A.4}$$

for the TT ranks.

It follows from the above considerations that

$$\mathcal{T}_{\leq \mathbf{r}} = \{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \mid \text{TT-}\operatorname{rank}(\mathbf{X}) \leq \mathbf{r}\},$$

[6] That is, 111, 211, 311, ..., 121, 221, 321, ...

where the inequality is understood elementwise. By the semicontinuity of matrix rank, this shows that $\mathcal{T}_{\leq \mathbf{r}}$ is closed.

Another useful description of the TT rank was given in [20]. If we look at the component tensors $\mathbf{U}_i$ of a $\mathrm{TT}_{\leq \mathbf{r}}$ decomposition $\mathbf{X} = \tau(\mathbf{U})$, we can unfold them either into the matrices

$$\mathbf{U}_i^L = \begin{bmatrix} U_i(1) \\ U_i(2) \\ \vdots \\ U_i(n_i) \end{bmatrix} \in \mathbb{R}^{(r_{i-1}n_i \times r_i)}, \tag{A.5}$$

or into

$$\mathbf{U}_i^R = \begin{bmatrix} U_i(1) & U_i(2) & \cdots & U_i(n_i) \end{bmatrix} \in \mathbb{R}^{r_{i-1} \times n_i r_i}.$$

These operations are called *left* and *right unfolding* of $\mathbf{U}_i$, respectively. With respect to the chosen ordering the following recursive relations hold:

$$\mathbf{U}^{\leq i} = (I_{n_i} \otimes \mathbf{U}^{\leq i-1})\mathbf{U}_i^L \tag{A.6}$$

and

$$\mathbf{U}^{\geq i} = \mathbf{U}_i^R(I_{n_i} \otimes \mathbf{U}^{\geq i+1}) \tag{A.7}$$

(where one should set $\mathbf{U}^{\leq 0} = \mathbf{U}^{\geq d+1} = 1$). Using (1.1) and (A.6) inductively, one can characterize the TT ranks as follows [20].

PROPOSITION A.2. *A tensor $\mathbf{X} = \tau(\mathbf{U})$ in the $TT_{\leq \mathbf{r}}$ format has TT rank $\mathbf{r}$ if and only if for $i = 1, 2, \ldots, d-1$ it holds that*

$$\operatorname{rank} \mathbf{U}_i^L = \operatorname{rank} \mathbf{U}_{i+1}^R = r_i. \tag{A.8}$$

Note that this implies the relations

$$r_i \leq r_{i-1}n_i \qquad \text{and} \qquad r_i \leq r_{i+1}n_{i+1}, \tag{A.9}$$

which, by induction, are equivalent to (A.4). The TT ranks r_i of a tensor $\mathbf{X}$ are hence not unrelated among each other. Since $r_0 = r_d = 1$, they can first increase for growing i but have to decrease from a certain index on.

Let us consider the set of tensors of fixed rank $\mathbf{r}$.

PROPOSITION A.3. *The set*

$$\mathcal{T}_{\mathbf{r}} = \{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \mid \text{TT-rank}(\mathbf{X}) = \mathbf{r}\} = \tau(\mathcal{U})$$

is either empty or a smooth an embedded submanifold in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ of dimension $\dim \mathcal{U} - \sum_{i=1}^{d-1} r_i^2$ and $\tau \colon \mathcal{U} \to \mathcal{T}_{\mathbf{r}}$ is a submersion, that is, $\tau|_{\mathcal{U}}$ is of constant rank $\dim \mathcal{U} - \sum_{i=1}^{d-1} r_i^2$.

Note that submersions are always open maps [7, paragraph (16.7.5)]. Hence τ maps open sets in $\mathcal{U}$ onto open sets in $\mathcal{T}_{\mathbf{r}}$. We do not give a proof of the above proposition here, but refer to the upcoming publication [43]. In the earlier work [20] it is shown that $\mathcal{T}_{\mathbf{r}}$ is an immersed submanifold.

We conclude with the remark that $\mathcal{T}_{\mathbf{r}}$ is not closed, but its closure is the set $\mathcal{T}_{\leq \mathbf{r}}$ of all $\mathrm{TT}_{\leq \mathbf{r}}$ decomposable tensors. This holds because, by the continuity of τ, the closure of $\mathcal{T}_{\mathbf{r}}$ contains $\tau(\overline{\mathcal{U}}) = \mathcal{T}_{\leq \mathbf{r}}$, which is already closed.

A.2. Equivalent TT representations. One difficulty of the $\mathrm{TT}_{\leq \mathbf{r}}$ format is that representations are highly nonunique. As stated in (1.2), we have $\tau(\mathbf{U}) = \tau(\hat{\mathbf{U}})$ whenever it holds that

(A.10)
$$\hat{U}_1(x_1) = U_1(x_1)A_1, \quad \hat{U}_d(x_d) = A_{d-1}^{-1}U_d(x_d), \quad \text{and} \quad \hat{U}_i(x_i) = A_{i-1}^{-1}U_i(x_i)A_i$$

for $i = 2, 3, \dots, d-1$, where the A_i are nonsingular $r_i \times r_i$ matrices. For rank-one tensors this is called *scaling indeterminacy* and we will use this terminology for TT tensors as well. The question whether the scaling indeterminacy (A.10) is the only kind of nonuniqueness has a simple answer.

PROPOSITION A.4. *A $TT_{\leq \mathbf{r}}$ decomposition $\mathbf{X} = \tau(\mathbf{U})$ is unique up to the scaling indeterminacy* (A.10) *if and only if $\mathbf{X}$ has TT rank $\mathbf{r}$.*

Proof. The necessity of this condition follows from the fact that a $\mathrm{TT}_{\leq \tilde{\mathbf{r}}}$ decomposition with $\tilde{r}_i \leq r_i$ can be artificially extended to a $\mathrm{TT}_{\leq \mathbf{r}}$ decomposition by adding zero blocks to the third-order components $\mathbf{U}_i$. This is then not covered by the operation (A.10). On the other hand, if $\operatorname{rank} \mathbf{X}^{(i)} = r_i$ for $i = 1, 2, \dots, d-1$, then (A.3) implies $\hat{\mathbf{U}}^{\leq i} = \mathbf{U}^{\leq i}A_i$ and $\hat{\mathbf{U}}^{\geq i+1} = A_i^{-1}\mathbf{U}^{\geq i+1}$ for nonsingular A_i. Using (A.6) one finds that this is equivalent to (A.10), which proves the sufficiency of the condition. ∎

The scaling operation (A.10) can be regarded as the action of the Lie group $\mathcal{G} = \times_{i=1}^{d-1} GL(r_i)$ on $\mathcal{U}$, which we denote by

$$\theta \colon \mathcal{U} \times \mathcal{G} \to \mathcal{U} \colon (\mathbf{U}, \mathsf{A}) \mapsto \hat{\mathbf{U}} =: \theta(\mathbf{U}, \mathsf{A}),$$

with $\hat{\mathbf{U}}$ defined by (A.10). Obviously θ is continuous. As in the main text, for fixed $\mathbf{U} \in \mathcal{U}$ we use the shorthand $\theta_{\mathbf{U}}$ for the map $\mathsf{A} \mapsto \theta(\mathbf{U}, \mathsf{A})$, and denote by

$$\mathcal{M}_{\mathbf{U}} = \{\hat{\mathbf{U}} = \theta_{\mathbf{U}}(\mathsf{A}) \mid \mathsf{A} \in \mathcal{G}\}$$

the orbits of the group action. Proposition A.4 states that

$$\mathcal{M}_{\mathbf{U}} = \{\hat{\mathbf{U}} \in \overline{\mathcal{U}} \mid \tau(\hat{\mathbf{U}}) = \tau(\mathbf{U})\}. \tag{A.11}$$

One can prove that $\mathcal{M}_{\mathbf{U}}$ is an embedded (not connected) submanifold of $\mathcal{U}$ of dimension $\dim \mathcal{G} = \sum_{i=1}^{d-1} r_i^2$. The tangent space of $\mathcal{M}_{\mathbf{U}}$ at $\mathbf{U}$ will be denoted by $T\mathcal{M}_{\mathbf{U}}$. Using the fact that the derivative of the matrix inverse $A \mapsto A^{-1}$ at a point A is the linear mapping $H \mapsto -A^{-1}HA^{-1}$, one calculates

$$\begin{aligned}\theta'_{\mathbf{U}}(\mathsf{A})[\mathsf{H}] = (&\mathbf{U}_1 H_1\,,\, -A_1^{-1}H_1A_1^{-1}\mathbf{U}_2A_2 + A_1^{-1}\mathbf{U}_2H_2, \dots \\ &\dots,\, -A_{d-2}^{-1}H_{d-2}A_{d-2}^{-1}\mathbf{U}_{d-1}A_{d-1} + A_{d-2}^{-1}\mathbf{U}_{d-1}H_{d-1}\,,\, -A_{d-1}^{-1}H_{d-1}A_{d-1}^{-1}\mathbf{U}_{d-1}),\end{aligned}$$

where $\mathsf{H} = (H_1, H_2, \dots, H_{d-1}) \in \bigtimes_{i=1}^{d-1} \mathbb{R}^{r_i \times r_i}$, and expressions like $A\mathbf{U}_iB$ with matrices A, B are understood slicewise as $AU_i(x_i)B$.[7] Evaluating this formula at the

[7] Denoting the inverse operations of the left and right as unfolding both by ten, it holds that $A\mathbf{U}_iB = \operatorname{ten}(A[\operatorname{ten}(\mathbf{U}_i^L B)]^R)$.

identity $\mathsf{I} = (I_{r_1}, I_{r_2}, \dots, I_{r_{d-1}})$ gives

$$T\mathcal{M}_{\mathbf{U}} = \left\{ \theta'_{\mathbf{U}}(\mathsf{I})[\mathsf{H}] \,\middle|\, \mathsf{H} \in \bigtimes_{i=1}^{d-1} \mathbb{R}^{r_i \times r_i} \right\}$$

$$= \Bigg\{ (\mathbf{U}_1 H_1, -H_1\mathbf{U}_2 + \mathbf{U}_2 H_2, \dots \tag{A.12}$$

$$\dots, -H_{d-2}\mathbf{U}_{d-1} + \mathbf{U}_{d-1}H_{d-1}, -H_{d-1}\mathbf{U}_{d-1}) \,\Big|\, \mathsf{H} \in \bigtimes_{i=1}^{d-1} \mathbb{R}^{r_i \times r_i} \Bigg\}.$$

A.3. Normalization. Elements in the same orbit parametrize the same TT tensor and are, therefore, called *equivalent*. It would be convenient and useful to find a normal form of TT rank $\mathbf{r}$ tensors which fixes a representation within the orbits. An ideal *normalization operator* $R\colon \mathcal{U} \to \mathcal{U}$ would have the properties (i) $R(\mathbf{U}) \in \mathcal{M}_{\mathbf{U}}$ and (ii) $R|_{\mathcal{M}_{\mathbf{U}}}$ is constant for all $\mathbf{U} \in \mathcal{U}$. Since $\mathcal{M}_{\mathbf{U}}$ is not connected, one might only require that $R|_{\mathcal{M}_{\mathbf{U}}}$ is constant on the connected components of $\mathcal{M}_{\mathbf{U}}$, but even then it is not clear whether such an operator can be continuously defined in a natural way. For our purposes, we need only a local variant. We use the shorthand $\mathsf{A}(\mathbf{U})$ for a function $\overline{\mathcal{U}} \to \mathcal{G}$.

DEFINITION A.5. *An operator of the form*

$$R\colon \mathcal{O} \subseteq \mathcal{U} \to \mathcal{U}\colon \mathbf{U} \mapsto \theta_{\mathbf{U}}(\mathsf{A}(\mathbf{U}))$$

defined on an open subset $\mathcal{O}$ of $\mathcal{U}$ is called local normalization operator *if*

(i) $R^2(\mathbf{U}) = R(\mathbf{U})$ *for all* $\mathbf{U} \in \mathcal{O}$,
(ii) R *is smooth in a neighborhood of its fixed points,*
(iii) $R|_{\mathcal{M}_{\mathbf{U}}}$ *is constant in a neighborhood of fixed points* $\mathbf{U}$.

The existence of local normalization operators is guaranteed.

PROPOSITION A.6. *For every $\mathbf{U}^* \in \mathcal{U}$ there exists an open neighborhood $\mathcal{O}$ of $\mathbf{U}^*$ and a local normalization operator $R_{\mathbf{U}^*}\colon \mathcal{O} \to \mathcal{U}$ such that $R_{\mathbf{U}^*}(\mathbf{U}^*) = \mathbf{U}^*$.*

Surely, this assertion can be pieced together from similar results in textbooks (see, e.g., [7, paragraph (16.10.3.2)], but one first has to verify that the Lie group action is proper, which is trivial here). The intuitive idea behind the claim is that the orbits $\mathcal{M}_{\mathbf{U}}$ completely fill out $\mathcal{U}$. By putting for fixed $\mathbf{U}^*$ a submanifold $\mathcal{N}_{\mathbf{U}^*}$ of co-dimension $\dim \mathcal{M}_{\mathbf{U}^*}$ through $\mathbf{U}^*$ and transversal to $\mathcal{M}_{\mathbf{U}^*}$, we can define $R_{\mathbf{U}^*}$ by the property $R_{\mathbf{U}^*}(\mathbf{V}) = \theta_{\mathbf{V}}(\mathsf{A}(\mathbf{V})) \in \mathcal{N}_{\mathbf{U}^*}$. This is illustrated in Figure A.1. For the most obvious choice $\mathcal{N}_{\mathbf{U}^*} = \mathbf{U}^* + W_{\mathbf{U}^*}$, where $W_{\mathbf{U}^*}$ is any complementary linear space to $T\mathcal{M}_{\mathbf{U}^*}$ (see, for instance, (3.1)), the existence of such a function $\mathsf{A}(\mathbf{V})$ for $\mathbf{V}$ in a neighborhood of $\mathbf{U}^*$ can be easily obtained from the implicit function theorem.

Alternatively, taking Proposition A.3 for granted, it follows from standard theorems (e.g., [7, paragraph (16.8.8)]) that a neighborhood $\mathcal{O}(\mathbf{U}^*)$ of $\mathbf{U}^*$ in $\mathcal{N}_{\mathbf{U}^*} = \mathbf{U}^* + W_{\mathbf{U}^*}$ is diffeomorphic to a neighborhood $\mathcal{O}(\mathbf{X})$ of $\mathbf{X}^* = \tau(\mathbf{U}^*)$ in $\mathcal{T}_{\mathbf{r}}$ via the mapping $\tau|_{\mathcal{N}_{\mathbf{U}^*}}$. One then can define $R_{\mathbf{U}^*} = (\tau|_{\mathcal{N}_{\mathbf{U}^*}})^{-1} \circ \tau$ on $\mathcal{O}(\mathbf{U}^*)$. Note that $R_{\mathbf{U}^*}$ then will be a local normalization operator for all $\mathbf{V} \in \tau|^{-1}_{\mathcal{N}_{\mathbf{U}^*}}(\mathcal{O}(\mathbf{X}))$. These $\mathbf{V}$ are the unique intersection points of $\mathbf{U}^* + W_{\mathbf{U}^*}$ with orbits close to $\mathbf{U}^*$. Conversely, one can show that for any $\mathbf{X}^* = \tau(\mathbf{U}^*) \in \mathcal{T}_{\mathbf{r}}$ and $\mathbf{X}$ close enough to $\mathbf{X}^*$, the corresponding neighborhood $\mathcal{O}(\mathbf{X})$ contains $\mathbf{X}^*$. In this way one can at least construct a local normalization operator for some $\mathbf{U}^* \in \mathcal{M}_{\mathbf{U}^*}$ without exactly knowing $\mathbf{U}^*$, which might be the unknown solution of a minimization problem as discussed later.

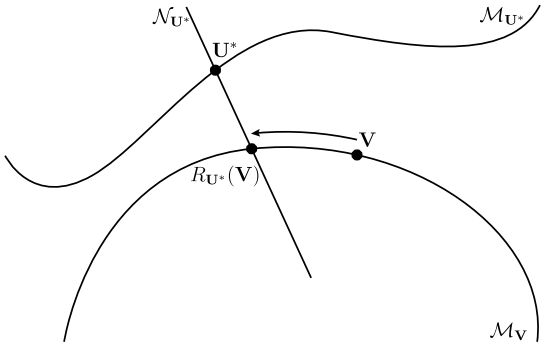

FIG. A.1. *The local normalization operator* $R_{\mathbf{U}^*}$.

Finding an explicit instance of a "self consistent" local normalization operator, which does not depend on a certain $\mathbf{U}^*$, is a nontrivial task. Usually, one uses (A.10) to choose a representation which satisfies

$$(\mathbf{U}_i^L)^T \mathbf{U}_i^L = I_{r_i} \tag{A.13}$$

for all i, which implies $(\mathbf{U}^{\le i})^T \mathbf{U}^{\le i} = I_{r_i}$. If $\mathbf{U} \in \mathcal{U}$, then such a representation is unique up to scaling by orthogonal matrices. We call a $\mathrm{TT}_{\le \mathbf{r}}$ representation satisfying (A.13) *left orthonormal*.

The successive TT-SVD algorithm of Oseledets [33] (also see [16, Chapter 12]), which decomposes a tensor of TT rank $\mathbf{r}$ into the TT format produces a certain normal form which is also used in the quantum physics community [45, 44]. This representation can be characterized by the property that the columns of $\mathbf{U}^{\le i}$ consist of left singular vectors of the ith unfolding $\mathbf{X}^{(i)} = \mathbf{U}^{\le i}\mathbf{U}^{\ge i+1}$, that is,

$$\mathbf{U}^{\ge i+1} = \Sigma_i V_i,$$

such that

$$\mathbf{X}^{(i)} = \mathbf{U}^{\le i}\Sigma_i V_i \tag{A.14}$$

is an SVD with singular values arranged in *descending* order. We call such a representation *left normal*. Note that such a representation is also right orthogonal in the sense that

$$\mathbf{U}^{\ge i+1}(\mathbf{U}^{\ge i+1})^T = \Sigma_i^2.$$

In the case that all unfoldings $\mathbf{X}^{(i)}$ have rank r_i and the nonzero singular vectors are mutually distinct (have multiplicity one), the left normal form of a $\mathrm{TT}_{\le \mathbf{r}}$ tensor is unique up to scaling by orthogonal diagonal matrices (distributions of signs). We will call tensors with this property *nondegenerate*.

It follows from (A.7) (and was already shown in [23]) that a left normal representation satisfies the two gauge conditions

$$(\mathbf{U}_i^L)^T \mathbf{U}_i^L = I_{r_i}, \quad \mathbf{U}_{i+1}^R (I_{n_{i+1}} \otimes \Sigma_{i+1}^2)(\mathbf{U}_{i+1}^R)^T = \Sigma_i^2, \tag{A.15}$$

of which the second one can be written pointwise as

$$U_{i+1}(x_j)\Sigma_{i+1}^2(U_{i+1}(x_j))^T = \Sigma_i^2, \quad j = 1, 2, \ldots, n_{i+1}. \tag{A.16}$$

On the other hand, if $\mathbf{X}$ is nondegenerate, a solution of (A.15) is unique up to scaling by orthogonal diagonal matrices, which shows that (A.15) is equivalent to (A.14).

We can define a so-called *left normalization operator*

$$R_L : \mathcal{U} \to \mathcal{U},$$

which scales a $\mathrm{TT}_{\leq \mathbf{r}}$ representation $\mathbf{U}$ into the left normal form, using, for instance, the successive SVD algorithm. In practice, that is, in the implementation, the stability and continuity of this procedure depends on related properties of the used SVD solver, which might be a black box. We therefore have to make some assumptions.

PROPOSITION A.7. *Assume that*

(i) *R_L is smooth in a neighborhood of its fixed points, and*
(ii) *left normal representations are fixed points of R_L.*

Then R_L is a local normalization operator in the neighborhood of left normal representations $\mathbf{U} \in \mathcal{U}$, if $\mathbf{X} = \tau(\mathbf{U})$ is nondegenerate.

Proof. Since, in the case of nondegeneracy, the left normal TT decomposition is unique up to scaling by orthogonal diagonal matrices, which form a discrete group, the fixed points of $R|_{\mathcal{M}_\mathbf{U}}$ are isolated points of its range. The assertion then follows from property (i) above. ∎

Similar statements can be made for the analogously defined *right orthonormal* and *right normal* representations.

REFERENCES

[1] J. BALLANI, L. GRASEDYCK, AND M. KLUGE, *Black box approximation of tensors in hierarchical Tucker format*, Linear Algebra Appl., 438 (2013), pp. 639–657.

[2] J. BALLANI AND L. GRASEDYCK, *A projection method to solve linear systems in tensor format*, Numer. Linear Algebra Appl., 20 (2013), pp. 27–43.

[3] J. C. BEZDEK AND R. J. HATHAWAY, *Convergence of alternating optimization*, Neural Parallel Sci. Comput., 11 (2003), pp. 351–368.

[4] D. CONTE AND C. LUBICH, *An error analysis of the multi-configuration time-dependent Hartree method of quantum dynamics*, M2AN Math. Model. Numer. Anal., 44 (2010), pp. 759–780.

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

[6] 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.

[7] J. DIEUDONNÉ, *Treatise on Analysis, Vol.* III, Academic Press, New York, 1972.

[8] L. ELDÉN AND B. SAVAS, *Perturbation theory and optimality conditions for the best multilinear rank approximation of a tensor*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 1422–1450.

[9] M. ESPIG, *Effziente Bestapproximation mittels Summen von Elementartensoren in hohen Dimensionen*, Ph.D. thesis, Universität Leipzig, Leipzig, Germany, 2007.

[10] M. ESPIG, W. HACKBUSCH, S. HANDSCHUH, AND R. SCHNEIDER, *Optimization problems in contracted tensor networks*, Comput. Vis. Sci., 14 (2011), pp. 271–285.

[11] M. ESPIG, W. HACKBUSCH, T. ROHWEDDER, AND R. SCHNEIDER, *Variational calculus with sums of elementary tensors of fixed rank*, Numer. Math. 122 (2012), pp. 469–488.

[12] A. FALCÓ AND W. HACKBUSCH, *On minimal subspaces in tensor representations*, Found. Comput. Math., 12 (2012), pp. 765–803.

[13] A. FALCÓ AND A. NOUY, *Proper generalized decomposition for nonlinear convex problems in tensor Banach spaces*, Numer. Math., 121 (2012), pp. 503–530.

[14] L. GRASEDYCK, *Hierarchical singular value decomposition of tensors*, SIAM. J. Matrix Anal. Appl., 31 (2010), pp. 2029–2054.

[15] L. GRASEDYCK AND W. HACKBUSCH, *An introduction to hierarchical ($\mathcal{H}$-) rank and TT-rank of tensors with examples*, Comput. Methods Appl. Math, 3 (2011), pp. 291–304.

[16] W. HACKBUSCH, *Tensor Spaces and Numerical Tensor Calculus*, Springer, Berlin, 2012.

[17] W. HACKBUSCH AND S. KÜHN, *A new scheme for the tensor representation*, J. Fourier Anal. Appl., 15 (2009), pp. 706–722.

[18] E. Hairer, C. Lubich, and G. Wanner, *Geometrical Numerical Integration—Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd ed., Springer, Berlin, 2006.
[19] C. J. Hillar and L.-H. Lim, *Most tensor problems are NP hard*, preprint, available at http://arxiv.org/abs/0911.1393.
[20] S. Holtz, T. Rohwedder, and R. Schneider, *On manifolds of tensors of fixed TT-rank*, Numer. Math., 120 (2012), pp. 701–731.
[21] 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.
[22] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[23] T. Huckle, K. Waldherr, and T. Schulte-Herbrüggen, *Exploiting matrix symmetries and physical symmetries in matrix product states and tensor trains*, Linear Multilinear Algebra, to appear; preprint available online at arXiv:1301.0746.
[24] 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.
[25] B. N. Khoromskij and I. V. Oseledets, *DMRG+QTT approach to computation of the ground state for the molecular Schrödinger operator*, preprint 69/2010, MPI MIS, Leipzig, Germany, 2010.
[26] O. Koch and C. Lubich, *Dynamical low-rank approximation of tensors*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2360–2375.
[27] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.
[28] D. Kressner and C. Tobler, *Preconditioned low-rank methods for high-dimensional elliptic PDE eigenvalue problems*, Comput. Methods Appl. Math., 11 (2011), pp. 363–381.
[29] 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
[30] C. Lubich, T. Rohwedder, R. Schneider, and B. Vandereycken, *Dynamical approximation by hierarchical Tucker and tensor-train tensors*, SIAM J. Matrix Anal. Appl., to appear.
[31] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[32] J. M. Ortega and M. L. Rockoff, *Nonlinear difference equations and Gauss–Seidel type iterative methods*, SIAM J. Numer. Anal., 3 (1966), pp. 497–513.
[33] I. V. Oseledets, *Tensor-train decomposition*, SIAM J. Sci. Comput., 33 (2011), pp. 2295–2317.
[34] I. V. Oseledets, *TT Toolbox* 2.0: *Fast multidimensional array operations in TT format*, preprint, available online at http://spring.inm.ras.ru/osel/wp-content/plugins/wp-publications-archive/openfile.php?action=open&file=27.
[35] I. V. Oseledets and S. V. Dolgov, *Solution of linear systems and matrix inversion in the TT-format*, SIAM J. Sci. Comput., 34 (2012), pp. A2718–A2739.
[36] I. V. Oseledets and E. E. Tyrtyshnikov, *TT-cross approximation for multidimensional arrays*, Linear Algebra Appl., 432 (2010), pp. 70–88.
[37] I. V. Oseledets and E. E. Tyrtyshnikov, *Tensor tree decomposition does not need a tree*, preprint 2009-08, INM RAS, Moscow, Russia, 2010.
[38] S. Schechter, *Iteration methods for nonlinear problems*, Trans. Amer. Math. Soc., 104 (1962), pp. 179–189.
[39] U. Schollwöck, *The density-matrix renormalization group*, Rev. Modern Phys., 77 (2005), pp. 259–315.
[40] U. Schollwöck, *The density-matrix renormalization group in the age of matrix product states*, Ann. Physics, 326 (2011), pp. 96–192.
[41] L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311.
[42] A. Uschmajew, *Local convergence of the alternating least squares algorithm for canonical tensor approximation*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 639–652.
[43] A. Uschmajew and B. Vandereycken, *The geometry of algorithms using hierarchical tensors*, Linear Algebra Appl., accepted; preprint available online at http://sma.epfl.ch/~uschmaje/papers/geom_htucker.pdf.
[44] F. Verstraete, J. I. Cirac, J. I. Latorre, E. Rico, and M. M. Wolf, *Renormalization-group transformations on quantum states*, Phys. Rev. Lett., 94 (2005), 140601.
[45] G. Vidal, *Efficient classical simulation of slightly entagled quantum computation*, Phys. Rev. Lett., 91 (2003), 147902.
[46] J. Weissinger, *Verallgemeinerungen des Seidelschen Iterationsverfahrens*, Z. Angew. Math. Mech., 33 (1953), pp. 155–163 (in German).

[47] S. White, *Density matrix formulation for quantum renormalization groups*, Phys. Rev. Lett., 69 (1992), pp. 2863–2866.
[48] S. White, *Density matrix renormalization group algorithms with a single center site*, Phys. Rev. B, 72 (2005), 180403.
[49] 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.