Numerische Mathematik (2019) 143:855–904
https://doi.org/10.1007/s00211-019-01072-4

Numerische Mathematik

Stable ALS approximation in the TT-format for rank-adaptive tensor completion

Lars Grasedyck[1] · **Sebastian Krämer**[1]

Received: 13 February 2018 / Revised: 28 March 2019 / Published online: 30 August 2019

Abstract
Low rank tensor completion is a highly ill-posed inverse problem, particularly when the data model is not accurate, and some sort of regularization is required in order to solve it. In this article we focus on the calibration of the data model. For alternating optimization, we observe that existing rank adaption methods do not enable a continuous transition between manifolds of different ranks. We denote this characteristic as *instability (under truncation)*. As a consequence of this property, arbitrarily small changes in the iterate can have arbitrarily large influence on the further reconstruction. We therefore introduce a singular value based regularization to the standard alternating least squares (ALS), which is motivated by averaging in microsteps. We prove its *stability* and derive a natural semi-implicit rank adaption strategy. We further prove that the standard ALS microsteps for completion problems are only stable on manifolds of fixed ranks, and only around points that have what we define as *internal tensor restricted isometry property, iTRIP*. In conclusion, numerical experiments are provided that show improvements of the reconstruction quality up to orders of magnitude in the new Stable ALS Approximation (SALSA) compared to standard ALS and the well known Riemannian optimization RTTC.

Mathematics Subject Classification 15A18 · 15A69 · 65F22 · 90C06 · 90C31

Both authors gratefully acknowledge support by the DFG priority programme 1648 under Grant GR3179/3-1.

✉ Sebastian Krämer
kraemer@igpm.rwth-aachen.de

Lars Grasedyck
lgr@igpm.rwth-aachen.de

[1] IGPM, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany

1 Introduction

Low rank tensor completion is a highly ill-posed inverse problem. In order for any recovery to succeed, regularity assumptions are required. This is either achieved by adding certain penalty terms, or by using an explicit reduction of degrees of freedom, which in the context of high-dimensional tensors can be obtained by using low rank representations, cf. [16,18,20]. The main goal of this work is to derive a rank adaptive method for tensor completion and moreover to discuss the benefits of such as well as the reasons why heuristics tend to be insufficient. Since the concept of *stability* in the sense of Definition 1 has not yet been considered in literature, the initial part is dedicated to the simpler matrix case (Sects. 2 and 3) in order to provide an easier access. The subsequent analysis will focus on the importance of these concepts to least squares tensor completion where the calibration of model complexity is more challenging.

1.1 Introduction to stability for ill-posed inverse problems through the example of matrix completion

In the setting of low rank matrix completion, the target of recovery is a matrix $M \in \mathbb{R}^{n\times m}$ which is only observable at points

$$M|_P = \{M_p\}_{p\in P} \in \mathbb{R}^P \cong \mathbb{R}^{|P|} \quad \text{for} \quad P \subset \mathcal{I} := \{1,\dots,n\}\times\{1,\dots,m\},$$

where P is a given, fixed sampling set, which we hence can not enlarge. One very strict regularity assumption is given by $\operatorname{rank}(M) = r$ for some sufficiently small $r \in \mathbb{N}$, which leads to the minimization problem

$$\begin{aligned} &\text{minimize } \|A - M\|_P \\ &\text{subject to } A \in \mathbb{R}^{n\times m},\ \operatorname{rank}(A) \le r, \end{aligned}$$

where $\|B\|_{\mathrm{P}}^2 := \sum_{i\in P} B_i^2$ for matrices B. A favorable data model for this task is the low rank representation, i.e. a function

$$\tau_r : (X,Y) \mapsto A = XY \in \mathbb{R}^{n\times m} \quad \text{for} \quad (X,Y) \in \mathcal{D}_r := \mathbb{R}^{n\times r}\times\mathbb{R}^{r\times m}.$$

Since every matrix has a unique rank, we can partition the target space $\mathbb{R}^{n\times m}$ into the disjoint subsets

$$\mathcal{T}_r := \{A \mid \operatorname{rank}(A) = r\}, \quad r = 0,\dots,\min(n,m).$$

With $\operatorname{image}(\tau_r) = \bigcup_{\tilde{r}\le r} \mathcal{T}_{\tilde{r}}$ in mind, the optimization is performed on the representation or data space $\mathcal{D}_r$. In an alternating least squares (ALS) method for example, one

then applies two optimization methods $\mathscr{M}^{(1)}, \mathscr{M}^{(2)}$,

$$\mathscr{M}_r^{(1)}(X, Y) := \left(\underset{\widetilde{X}}{\operatorname{argmin}} \, \|\widetilde{X}Y - M\|_P, Y \right), \tag{1}$$

$$\mathscr{M}_r^{(2)}(X, Y) := \left(X, \underset{\widetilde{Y}}{\operatorname{argmin}} \, \|X\widetilde{Y} - M\|_P \right), \tag{2}$$

which in this context are called microsteps. Note that the argmin is not necessarily unique, and we choose the element minimizing the Frobenius norm $\|XY\|_F$. Formally, for each value of the matrix rank r, every single $\mathscr{M}_r^{(1)}, \mathscr{M}_r^{(2)}$ is a different function.

For most realistic applications, it is more reasonable to relax the regularity assumption to M being *nearly* rank r, which means that after r entries, the singular values of M become sufficiently smaller. Ultimately, if no assumptions are made, the appropriate model complexity is a matter of the quality and magnitude of P with respect to M. Yet in the general case, the missing structure of given data hardly allows to obtain knowledge about this relation. Therefor, since overestimating the model complexity (i.e. the rank) ultimately leads to flawed results, a cautious learning process is required to adapt such, as rank increasing strategies already suggest. Due to the difficult nature of the problem, we do not expect to be able to find the global minimizer, but instead focus on single aspects that are likely to improve the approximation quality.

Each adaption of the rank during the optimization will cause the algorithm to change between data spaces $\mathscr{D}_r$. Intuitively, considering that the generated spaces $\mathscr{T}_r$ have pairwise distance 0 within $\mathbb{R}^{n\times m}$, one would want that a change of rank does not have large impact, given the problematic nature of overfitting. This, however, is not true for both $\mathscr{M}^{(1)}, \mathscr{M}^{(2)}$, while arbitrarily small perturbations of the iterate $A = \tau_r(X, Y)$ may change its rank. For ill-posed inverse problems, we hence propose the following concept:

Definition 1 (*Stability*) Let $\mathscr{M}$ be a method that maps any rank r to a function $\mathscr{M}_r : \mathscr{D}_r \to \mathscr{D}_r$ (the optimization method for fixed rank). We define the following properties (Fig. 1):

- $\mathscr{M}$ is called *representation independent*, if $\tau_r(\mathscr{M}_r(G)) = \tau_r(\mathscr{M}_r(\widetilde{G}))$ for all r and $G, \widetilde{G} \in \mathscr{D}_r$ with $\tau_r(G) = \tau_r(\widetilde{G})$. We then define τ_r^{-1} to map to one possible representation (we want to circumvent the use of equivalence classes).
- $\mathscr{M}$ is called *fixed-rank stable*, if it is representation independent and for any fixed rank r, the map $\tau_r \circ \mathscr{M}_r \circ \tau_r^{-1} : \mathscr{T}_r \to \mathbb{R}^{\mathscr{I}}$ is continuous.
- $\mathscr{M}$ is called *stable*, if it is representation independent and the function

$$f_{\mathscr{M}} : \mathbb{R}^{\mathscr{I}} \to \mathbb{R}^{\mathscr{I}}, \quad f_{\mathscr{M}}(A) := \tau_{r(A)} \circ \mathscr{M}_{r(A)} \circ \tau_{r(A)}^{-1}(A), \tag{3}$$

  where $r(A)$ is the rank of A, is continuous.

This definition of stability is not restricted to matrix completion. Data spaces, the rank (possibly generalized to any model complexity) and the method can be replaced by any appropriate type, in particular tensor completion in hierarchical tensor formats (cf.

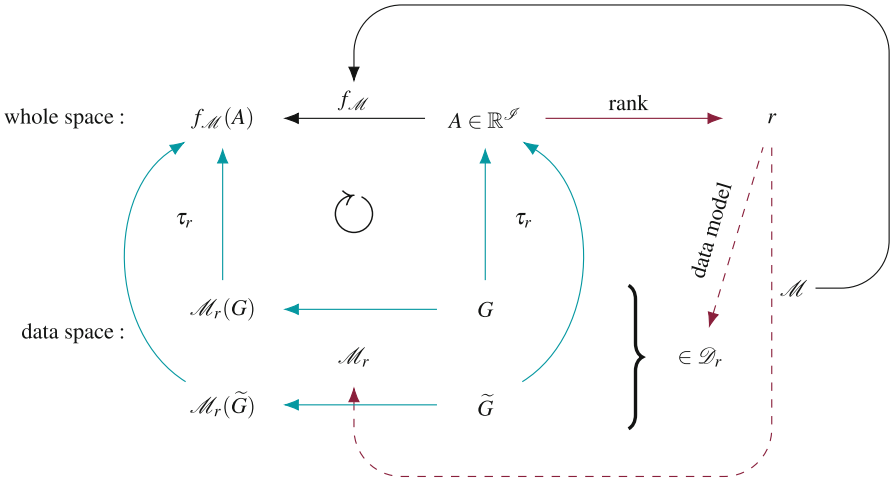

Fig. 1 The diagram depicting Definition 1. Magenta part: depending on the rank of A, the method $\mathscr{M}$ provides a specific mapping $\mathscr{M}_r$ to be applied to equivalent representations $G, \widetilde{G} \in \mathscr{D}_r$. Teal part: representation independent states that $f_{\mathscr{M}}$ is well-defined since both lower paths from A along the data space result in the same output within the whole space. Stability requires that this function, the upper path, is continuous (color figure online)

Sect. 4). The only assumption should usually be a nestedness of spaces, i.e. $\mathscr{T}_r \subset \mathscr{T}_{\widetilde{r}}$ whenever $r \preceq \widetilde{r}$ (entry-wise inequality).

Properly calibrating the rank r for unstable methods in the context of ill-posed inverse problems may lead to complications. Most of the operators applied to representations are stable, e.g. truncations based on matrix singular values. The situation however changes if we apply the partial optimization (or micro-) step $\mathscr{M}^{(1)}$ or $\mathscr{M}^{(2)}$ on a low rank representation:

Example 1 (*Instability of alternating least squares matrix completion steps*) Let $a \in \mathbb{R} \setminus \{0, 1\}$ be a possibly very small parameter. We consider the target matrix M and an ε-dependent initial approximation $A = A(\varepsilon)$

$$M := \begin{pmatrix} \boxed{?} & 1.1 & 0.9 \\ 1 & 1 & 1.1 \\ 1.1 & 1 & 1 \end{pmatrix}, \quad A(\varepsilon) := \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} + \varepsilon \begin{pmatrix} 0.5+a & 0.5+a & -a \\ 1+a & 1+a & -1-a \\ 1-a & 1-a & -1+a \end{pmatrix},$$

where the entry $M_{1,1}$ (the question mark above) is not known or given. The matrix M is of rank 3 and $A(\varepsilon)$ is of rank $r = 1$ for $\varepsilon = 0$ and of rank $r = 2$ otherwise. We seek a best approximation of (at most) rank 2 in the least squares sense for the known entries of M. In a single ALS step, as defined by (2), we replace $Y(\varepsilon)$ of the low rank representation $A(\varepsilon) = X(\varepsilon)Y(\varepsilon)$ by the local minimizer, where in this case

$$A(0) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}, \qquad A(\varepsilon) = \begin{pmatrix} 1 & 0.5+a \\ 1 & 1+a \\ 1 & 1-a \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & \varepsilon & -\varepsilon \end{pmatrix} \text{ if } \varepsilon > 0.$$

This optimization yields a new matrix, $B(\varepsilon) = f_{\mathscr{M}^{(2)}}(A(\varepsilon)) = \tau_r \circ \mathscr{M}_r^{(2)} \circ \tau_r^{-1}(A(\varepsilon))$ (independently of the chosen representation), given by

$$B(0) = \begin{pmatrix} 1.05 & * & * \\ 1.05 & * & * \\ 1.05 & * & * \end{pmatrix}, \qquad B(\varepsilon) = \begin{pmatrix} 1 + \frac{1}{40a} & * & * \\ 1.0 & * & * \\ 1.1 & * & * \end{pmatrix} \text{ if } \varepsilon > 0. \quad (* \text{ is some value})$$

Now let a be fixed and let ε tend to zero so that the initial guess $A(\varepsilon) \to A(0)$. However, $B(\varepsilon) \nrightarrow B(0)$, thus violating the stability. Furthermore, the rank two approximation $B(\varepsilon)$, given an arbitrary, fixed $\varepsilon > 0$, diverges as $a \to 0$, in particular it is not convergent although the initial guess $A(\varepsilon)$ converges to a rank two matrix as $a \to 0$. Thus, the microstep is not even stable for fixed rank. We want to stress that the initial guess is bounded for all $\varepsilon, a \in (0, 1)$, but the difference between $B(0)$ and $B(\varepsilon)$ is unbounded for $a \to 0$ (cf. Definition 11). The unboundedness can be remedied by adding a regularization term in the least squares functional, e.g. $+\|XY\|$, but the ALS step remains unstable.

This example likewise demonstrates that ALS for tensor completion is not stable and thus, as discussed before, problematic when adapting the rank (cf. Sect. 4). We will further show that this is not a marginal phenomenon, but occurs systematically during any rank change (cf. Example 2).

1.2 Relation to other matrix and tensor methods

Whenever a tensor is point-wise available, algorithms such as the TT-SVD [33] can just establish the exact rank based on its very definition or a reliable rank estimate as well as representation can be obtained through cross-approximation methods, a setting in which the subset of used entries can be chosen freely [4,32].

If only indirectly given, adapting the rank of the sought low rank tensor can still be straight-forward, e.g. when the rank has to be limited *only* due to computational complexity, while in principle the exact solution is desired [2,5]. Here, an optimal regulation of thresholding parameters becomes most important. This mainly includes classical problems that have been transferred to large scales. These may for example be solved with iterative methods [1,3,29], which naturally increase the rank and rely on subsequent reductions, or also by rank preservative optimization, such as alternating optimization [10,12,22,37], possibly combined with a separate rank adaption.

Provided that the tensor restricted isometry property holds, the task may be interpreted as distance minimization with respect to a norm that is sufficiently similar to the Frobenius norm and analyzed based on compressed sensing [35]. Black box tensor completion for a fixed sampling set, however, requires a certain solution to a positive-semi definite linear system. Hence neither an exact solution is reasonable nor does any norm equivalence hold. Thus, the available data is easily misinterpreted, the more so if the rank is overestimated, and truncation based algorithms, including DMRG [22,25], are misled.

Nuclear norm minimization, being closely related to compressed sensing as well, has a very strong theoretical background [7,8,17,36] for the matrix case. These

approaches rely on a direct adaption of the target function, that is convex relaxation. Yet it appears that they are outperformed in practice by alternating least squares approaches [24] and the simplifications required for an adaption to tensors [13,28,38] do not seem to allow for an appropriate generalization [31]. Also the approaches which themselves retreat to alternating least squares [21] treat the iterations as necessity for the minimization of an objective function with regularization term. The penalty term is, as usual, based on the singular values of the output of a microstep (a posteriori), as it is also the case for the work [40] on tensor completion through Riemannian optimization. Although their term may appear similar to the term we derive (cf. Theorem 1), the stability property, on the contrary, requires that the penalty term depends on the current singular values before the microstep (a priori), which is an essential difference: we explicitly allow and exploit small singular values instead of penalizing such. In that sense, we treat each update and adaption as part of a *learning progress*, where the magnitudes of singular values indicate in some respects an uncertainty of approximation. To the best of our knowledge, this point of view has not yet been considered in literature and hence a concept of stability as we define it has not been investigated.

For fixed or uniform rank, there have been least squares based proposals in hierarchical tensor formats [27,39] as well. The essential adaption of the rank however, including similar matrix approaches, is rarely considered, all the less in numerical tests, and remains an open problem in this setting. A mentionable approach so far is the rank increasing strategy [15,43] and its regularization properties are a first starting point for this article.

The rest of the article is organized as follows: In Sect. 2, we further investigate instability and exemplarily analyze approaches towards it in the matrix case. Based on this insight, we motivate a variational residual function, derive its minimizer and present a stable algorithm for matrix completion in Sect. 3. In Sect. 4, we begin to generalize former results to high dimensional tensors, yet essentially work in three dimensions. In the main Sect. 5, we then derive its minimizer and prove stability (Theorem 4) for the thereby obtained regularized microsteps, further analyzing these in Sect. 6. Subsequently, in Sect. 7, these results are transferred back to arbitrarily dimensional tensors. Section 8 finishes with the necessary details for the algorithm, including its rank adaption as it is naturally given through stable alternating least squares. Comprehensive numerical tests (exclusively for unknown ranks) are provided in Sect. 9. Detailed tables with the numerical values shown in figures can be found in the "Appendix".

2 Instability and approaches to resolve the problem

The following example shows that instability can be observed systematically during rank changes in ALS, or more general, in any such range based optimization. In that sense, the implied complications for ill-posed, inverse problems may frequently occur:

Example 2 *(ALS for ill-posed, inverse problems is unstable)* Consider the microstep $\mathscr{M}^{(2)}$ as in (2). Let $U \in \mathbb{R}^{n\times r}$, $V \in \mathbb{R}^{m\times r}$ be orthogonal, such that $U\Sigma V^T$ is a truncated SVD of a rank r matrix $A = \tau_r(U, \Sigma V^T) \in \mathbb{R}^{n\times m}$. We now let $\sigma_r \to 0$, $\sigma_r > 0$, such that in the limit $A^* := A|_{\sigma_r=0}$ has rank $r-1$. The update is independent

of this last singular value though:

$$f_{\mathscr{M}^{(2)}}(A) = \tau_r\left(\mathscr{M}_r^{(2)}(U, \Sigma V^T)\right) = U \operatorname*{argmin}_Y \|UY - M\|_P = \lim_{\varepsilon \searrow 0} f_{\mathscr{M}^{(2)}}(A|_{\sigma_r=\varepsilon}) \tag{4}$$

However, if $\sigma_r = 0$, then $A|_{\sigma_r=0}$ has rank $r-1$ and a truncated SVD $U_c \Sigma_c V_c^T$. Hence, the update

$$f_{\mathscr{M}^{(2)}}(A|_{\sigma_r=0}) = \tau_{r-1}\left(\mathscr{M}_{r-1}^{(2)}\left(U_c, \Sigma_c V_c^T\right)\right) = U_c \operatorname*{argmin}_Y \|U_c Y - M\|_P$$

is in general different from the limit (4), given that the range of U is different from U_c. The same holds for an analogous update $\mathscr{M}^{(1)}$ of $X = U\Sigma$. Note that these updates are indeed representation independent.

The microsteps of ALS in the tensor case behave in the same way. The only difference is that there are two tuples of singular values $\sigma^{(\mu-1)}$ and $\sigma^{(\mu)}$ adjacent to a core G_μ (cf. Lemma 6). There may be many ways to stabilize the microsteps. However, we aim for an as little distorting as practical way to do this. A quite successful approach for completion has been the rank increasing strategy, e.g. [43]. Therein, the model complexity is slowly increased, step by step attempting a more distinct approximation. Thereby, local minima that correspond to overfitting are avoided.

A similar kind of effect can be achieved by assuming an uncertainty on the current iterate, or, equivalently, averaging the tensor update function. That way, the level of regularization can be adapted continuously and is less dependent on the discrete rank but the more meaningful, real-valued singular values. We will first view this in a minimal fashion for the matrix case and the method $\mathscr{M}^{(2)}$ defined by (2) (for the remainder of this section called $\mathscr{M}$ for simplicity). With this approach, we can motivate an algorithm that is stable under truncation and allows to straightforwardly adapt ranks. While other methods in fact keep distance from the border of one manifold of a fixed rank, we aim to optimize, in a certain sense, continuously between manifolds of different ranks.

Assuming local integrability of $f_{\mathscr{M}}$ [as defined in (3)], we obtain that the variational function

$$\begin{aligned} f^*_{\mathscr{M}}(A) &:= \frac{1}{|\mathbb{V}_{A,\omega}|} \int_{\mathbb{V}_{A,\omega}} f_{\mathscr{M}}(H) \mathrm{d}H \\ \mathbb{V}_{A,\omega} &:= \{H \in image(\tau_{\widetilde{r}}) \mid \|H - A\|_F \leq \omega\} \end{aligned} \tag{5}$$

is continuous within $image(\tau_{\widetilde{r}})$, where $\|\cdot\|_F$ is the Frobenius norm and $\widetilde{r}$ may be considered an upper bound to the rank (cf. Fig. 2). However, this function does not preserve low rank structure, apart from appearing to be to complicated to evaluate, and therefore we cannot find a method $\mathscr{M}^*$ for which $f^*_{\mathscr{M}} = f_{\mathscr{M}^*}$. Still, all simplifications which we will make remain subject to this motivation.

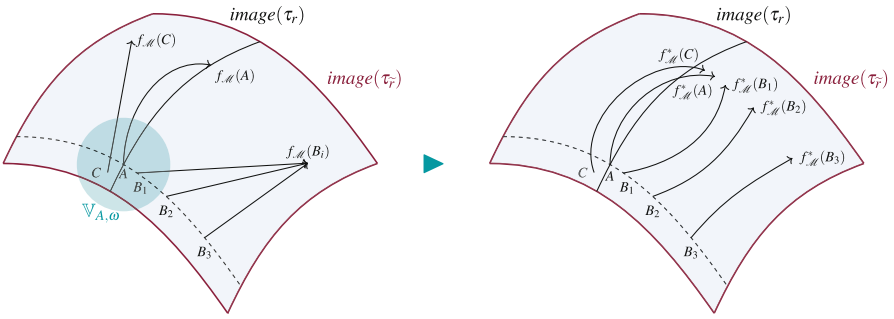

Fig. 2 The schematic display of the unstable function $f_{\mathscr{M}}$ (left) and the variational, stable $f^*_{\mathscr{M}}$ (right). In both pictures, the image of τ_r is depicted as black curve contained in the image of $\tau_{\widetilde{r}}$ shown as blue area (with magenta boundary). A is a rank r element, while C and each B_i has rank $\widetilde{r}$. Left: Regardless of their distance to A, the tensors B_1, B_2 and B_3 (and any other point of the dotted line except the lower rank element A) are mapped to the same point $f_{\mathscr{M}}(B_i)$. Likewise, C is, although as close to A as B_1, mapped to a completely different point. The teal circle exemplarily shows one possible range of averaging at the point A. Right: If an element (such as B_1 and C) is close to A, then this also holds for their function values. However, $f^*_{\mathscr{M}}(A)$ is not rank r anymore (in fact, the image of $f^*_{\mathscr{M}}$ is generally not even rank $\widetilde{r}$) (color figure online)

2.1 Investigations into the connection between averaging and stability

Although the following examples are heavily narrowed down, we observe which kind of simplifications we may apply in order to obtain a feasible, but still stable method (cf. Sect. 3). First, we consider a scenario in which we limit the perturbation that the left singular vectors U receive due to the variation of A to only one component as it is approximately the case if $\sigma_1 \gg \sigma_2 \approx \omega$.

Lemma 1 (Variational low rank matrix approximation) *Let $\mathscr{M}$ be defined by (2) and $P = \mathscr{I}$ (full sampling). Let further $A = U\Sigma V^T \in \mathbb{R}^{n\times m}$ be of rank two, given by its SVD components $U = (u_1 \mid u_2)$, $\Sigma = diag(\sigma_1, \sigma_2)$ and V as well as $M \in \mathbb{R}^{m\times m}$ arbitrary and $0 < \omega < \sqrt{2}\sigma_2$. Then*

$$\begin{aligned}\widehat{f}_{\mathscr{M}}(A) &:= \frac{1}{|V_\omega|}\int_{V_\omega} f_{\mathscr{M}}\left((u_1 \mid u_2 + \Delta u_2)\Sigma V^T\right)\mathrm{d}\Delta u_2 \\ &= \underbrace{u_1u_1^TM}_{\text{optimization}} + \underbrace{(1-\alpha_\omega)^2u_2u_2^TM}_{\text{regularization}} + \underbrace{\frac{2\alpha_\omega-\alpha_\omega^2}{m-2}(I_m - u_1u_1^T - u_2u_2^T)M}_{\text{replenishment}} \\ V_\omega &:= \Big\{\Delta u_2 \mid \|(u_1 \mid u_2 + \Delta u_2)\Sigma V^T - A\|_F = \omega, \\ &\qquad (u_1 \mid u_2 + \Delta u_2) \text{ has orthonormal columns}\Big\} \end{aligned} \tag{6}$$

for $\alpha_\omega = \frac{\omega^2}{2\sigma_2^2}$, such that $\alpha_\omega \to 1$ if $\omega \to \sqrt{2}\sigma_2$. Alternatively, considering complete uncertainty concerning the second singular vector, we obtain

$$\frac{1}{|V_\omega|}\int_{V_\omega} f_{\mathscr{M}}\left((u_1 \mid \Delta u_2)\Sigma V^T\right)\mathrm{d}\Delta u_2 = u_1u_1^TM + \frac{1}{m-1}(I_m - u_1u_1^T)M,$$

where here $V_\omega := \{\Delta u_2 \mid (u_1 \mid \Delta u_2)$ *has orthonormal columns*$\}$.

Proof We parameterize V_ω. First, $\omega = \|(u_1 \mid u_2 + \Delta u_2)\Sigma V^T - A\|_F = \|\Delta u_2\|_F\, \sigma_2$ and hence $\|\Delta u_2\|_F = \frac{\omega}{\sigma_2}$. By orthogonality conditions, we obtain $\Delta u_2 = -\alpha_\omega u_2 + \Delta u_2^\perp$ with $\Delta u_2^\perp \perp \mathrm{range}(U)$ for a fixed $\alpha_\omega = \frac{\omega^2}{2\sigma_2^2}$. Hence, V_ω is an $(m-3)$-sphere of radius $\beta_\omega = \sqrt{\frac{\omega^2}{\sigma_2^2} - \alpha_\omega^2}$, that is $\beta_\omega S^{m-2}$. The update for each instance of $\Delta u_2^\perp$ is given by

$$f_{\mathscr{M}}((u_1 \mid u_2 + \Delta u_2)\Sigma V^T) = (u_1 \mid u_2 + \Delta u_2)(u_1 \mid u_2 + \Delta u_2)^T M.$$

We integrate this over V_ω and obtain

$$\int_{V_\omega} f_{\mathscr{M}} = \int_{V_\omega} u_1 u_1^T M + \int_{V_\omega} (1-\alpha_\omega)^2 u_2 u_2^T M + \int_{V_\omega} \Delta u_2^\perp \Delta u_2^{\perp T} M$$

since all integrals of summands which contain $\Delta u_2^\perp$ exactly once vanish due to symmetry. We can simplify the last summand with Lemma 4 to

$$\int_{V_\omega} \Delta u_2^\perp \Delta u_2^{\perp T} M = \int_{\beta_\omega S^{m-2}} (Hx)(Hx)^T M \mathrm{d}x = H H^T \frac{2\alpha_\omega - \alpha_\omega^2}{m-2} |V_\omega| M$$

for a linear, orthonormal map H that maps $x \in \beta_\omega S^{m-2}$ to $\Delta u_2^\perp$, that is, embeds it into $\mathbb{R}^m$. One can then conclude that $HH^T = I_m - u_1 u_1^T - u_2 u_2^T$, since the rank of H is $m-2$ and $range(H) \perp range(U)$. The division by $|V_\omega|$ then finishes the first part. The second part is analogous. □

We can observe that, in this case, choosing ω close to σ_2, or in that sense a small σ_2, will *filter* out influence of u_2. This is indeed in agreement to the update which the rank 1 best-approximation to A would yield.[1] More importantly, the result $\widehat{f}_{\mathscr{M}}(A)$ in (6) is not low rank, yet close to the rank 2 approximation $U(u_1^T M \mid (1-\alpha_\omega)^2 u_2^T M)$, in which the first component U has remained the same. While the variational model as in (5) remains the basic idea, it is too complicated to be used for the derivation of a stable method $\mathscr{M}^*$. We instead consider a slightly modified approach (which will be used in the following Sects. 3 and 5).

Lemma 2 (Low rank matrix approximation using a variational residual function) *In the situation of Lemma 1, we have*

$$\underset{\widetilde{V}}{argmin} \frac{1}{|V_\omega|} \int_{V_\omega} \|(u_1 \mid u_2 + \Delta u_2)\widetilde{V} - M\|_F^2 \mathrm{d}\Delta u_2 = \left(u_1^T M \mid (1-\alpha_\omega) u_2^T M\right) \quad (7)$$

[1] Note that we fixed $\|\Delta u_2\|_F = \omega$ for simplicity as well as that for $\omega > \sqrt{2}\sigma$, Example 1 does not make sense. Allowing perturbations *up to* a magnitude ω will prohibit that the influence of u_2 vanishes completely, hence u_2 is never actually truncated.

Proof Let $\widetilde{V}$ be the corresponding minimizer. With the same derivation as in Lemma 1, we obtain

$$\begin{aligned} |V_\omega|\widetilde{V} &= \int_{V_\omega} (u_1 \mid u_2 + \Delta u_2)^T M \mathrm{d}\Delta u_2 \\ &= \left(u_1^T M |V_\omega| \mid u_2^T M |V_\omega| + \int_{V_\omega} -\alpha_\omega u_2^T M + \Delta u_2^\perp M \mathrm{d}\Delta u_2 \right) \\ &= |V_\omega| \left(u_1^T M \mid (1-\alpha_\omega) u_2^T M \right). \end{aligned}$$

□

Comparing this to the rank 2 approximation of the previous result (6), we observe that solely $(1-\alpha_\omega)^2$ has been replaced by $1-\alpha_\omega$. For our purpose, these terms are sufficiently similar for small $\alpha_\omega \geq 0$.

Remark 1 *(Replenishment and lower limit)* The so called replenishment term in (6) points at an important aspect which we analyze in Sect. 6. We later bypass related problems through an additional manipulation of singular values, the intensity of which is proportional to the current residual.

We here refer to a Matlab implementation of a (superficially random) Monte Carlo approach to the unsimplified variational microstep $f^*_{\mathscr{M}}$ as in (5) for matrix completion, which is linked on the personal webpage of the author Sebastian Krämer.[2] Likewise, an implementation of the final algorithm *SALSA* (Algorithms 2 and 4), which is developed from the idea in Lemma 2, can be found for the matrix case as well as for the tensor case.

3 Stable alternating least squares microsteps for matrix completion

In this section, we adapt the target function of each microstep $\mathscr{M}$ in order to obtain a stable method $\mathscr{M}^*$. When performing ALS, instead of just one specific iteration point $A \in \mathscr{T}_r$, the results in Sect. 2.1 suggest to instead consider a set $V_\omega(A)$ of variations, or uncertainty, ΔA along the manifold $\mathscr{T}_r$ of rank r matrices:

$$V_\omega(A) := \{\Delta A \mid A + \Delta A \in \mathscr{T}_r,\ \|\Delta A\|_F \leq \omega\}, \quad r = \operatorname{rank}(A)$$

The initial idea is slightly similar to gradient sampling (e.g. [6]), but is then pursued differently. Let $A = \tau_r(X, Y)$ and $(\Delta X, \Delta Y)$ such that $A + \Delta A = \tau_r(X + \omega\Delta X, Y + \omega\Delta Y)$. Then

$$\begin{aligned} \|\Delta A\|_F^2 &= \|(X + \omega\Delta X)(Y + \omega\Delta Y) - XY\|_F^2 \\ &= \|\omega(\Delta X Y + X \Delta Y)\|_F^2 + \left(\mathscr{O}(\omega^2)\right)^2 \end{aligned} \tag{8}$$

[2] By the time the paper is written, the address is www.igpm.rwth-aachen.de/team/kraemer.

The term $\|\Delta X Y + X \Delta Y\|_F^2$ can be approximated, assuming the angles between the three summands are small, by $\|\Delta X Y\|_F^2 + \|X \Delta Y\|_F^2$. This and Lemma 2 then motivate the following definition.

Definition 2 (*Variational residual function*) Let $\omega \geq 0$, M the target matrix, P the sampling set and $A = XY$ the current iterate. We define the variational residual function $C := C_{M,P,X,Y} : \mathscr{D}_r \to \mathbb{R}$ by

$$C(\widetilde{X}, \widetilde{Y}) := \int_{\mathbb{V}_\omega(X,Y)} \|(\widetilde{X} + \Delta X)(\widetilde{Y} + \Delta Y) - M\|_P^2 \mathrm{d}\Delta X \mathrm{d}\Delta Y,$$
$$\mathbb{V}_\omega(X, Y) := \{(\Delta X, \Delta Y) \mid \|\Delta X Y\|_F^2 + \|X \Delta Y\|_F^2 \leq \omega^2\}. \tag{9}$$

3.1 Minimizer of the variational residual function for matrices

We define our modified methods as

$$\left(\mathscr{M}^{(1)}\right)^*(X,\ Y) := \left(\operatorname*{argmin}_{\widetilde{X}} C(\widetilde{X}, Y),\ Y\right) \tag{10}$$

$$\left(\mathscr{M}^{(2)}\right)^*(X,\ Y) := \left(X,\ \operatorname*{argmin}_{\widetilde{Y}} C(X, \widetilde{Y})\right) \tag{11}$$

with $C = C_{M,P,X,Y}$ as in (9). It should further be noted that $\mathbb{V}_\omega$ does not depend on the unknown $\widetilde{X}$, $\widetilde{Y}$, respectively, but on the current iterate. We will later see that the minimizers are unique, but for formality we again use the minimization of the Frobenius norm of the iterate as secondary criterion.

Lemma 3 *The two methods* $(\mathscr{M}^{(1)})^*$ (10) *and* $(\mathscr{M}^{(2)})^*$ (11) *are representation independent.*

Proof We later prove this for the generalized tensor case (cf. Lemma 7). □

Here, and throughout the remainder of the paper, we use the following, convenient notations, since we often have to reshape, restrict or project objects.

Definition 3 (*Restrictions*) For any object $A \in \mathbb{R}^I$ and index set $S \subset I$, we use $A|_S \in \mathbb{R}^S$ as restriction. For a matrix M, let $M_{:,i}$ be its i-th column and $M_{i,:}$ be its i-th row. Furthermore, whenever we apply a restriction to an object or reshape it, we also use the same notation to correspondingly modify index sets (cf. Theorem 1).

Lemma 4 (Integral over all variations) *Let* $n, m \in \mathbb{N}$, $\omega \geq 0$ *and* $H \in \mathbb{R}^{n \times n}$ *be a matrix as well as*

$$V_\omega^{(n,m)} := \{X \in \mathbb{R}^{n \times m} \mid \|X\|_F = \omega\}.$$

Then

$$\int_{V_\omega^{(n,m)}} X^T H X \mathrm{d}X = \frac{\omega^2 |V_\omega^{(n,m)}|}{nm} trace(H) I_m, \quad \left|V_\omega^{(n,m)}\right| := \int_{V_\omega^{(n,m)}} 1.$$

Proof Let Y be the result of the above integral. Then

$$Y_{ij} = \operatorname{trace}(Y_{ij}) = \int_{V_\omega^{(n,m)}} \operatorname{trace}\big(X_{:,i}^T H X_{:,j}\big) \mathrm{d}X = \operatorname{trace}\left(H \int_{V_\omega^{(n,m)}} X_{:,j} X_{:,i}^T \right) \mathrm{d}X.$$

Due to symmetry, for some $\alpha \in \mathbb{R}$, we have

$$\int_{V_\omega^{(n,m)}} \operatorname{vec}(X)\operatorname{vec}(X)^T \mathrm{d}X = \alpha I_{nm}, \quad \int_{V_\omega^{(n,m)}} \operatorname{vec}(X)^T \operatorname{vec}(X) \mathrm{d}X = \omega^2 |V_\omega^{(n,m)}|.$$

Since the second term is the trace of the first one, it follows that $\alpha = \omega^2 |V_\omega^{(n,m)}|/(nm)$. We can hence simplify

$$Y_{ij} = \begin{cases} \omega^2 |V_\omega^{(n,m)}|/(nm) \operatorname{trace}(H) & \text{if } i = j, \\ 0 & \text{otherwise}, \end{cases}$$

which is the to be proven statement. □

We now derive the minimizer of the variational residual function for matrices (9). One can use an SVD of the current iterate $A = U \Sigma V^T$ for simplification. In this case, $\mathbb{V}_\omega$ as in Definition 2 takes the easier forms

$$\begin{aligned} \mathbb{V}_\omega\big(U\ \Sigma,\ V^T\big) &= \big\{(\Delta X, \Delta Y) \mid \big\|\Delta X V^T\big\|_F^2 + \|U \Sigma \Delta Y\|_F^2 \le \omega^2\big\} \\ &= \big\{(\Delta X, \Delta Y) \mid \|\Delta X\|_F^2 + \|\Sigma \Delta Y\|_F^2 \le \omega^2\big\}, \\ \mathbb{V}_\omega\big(U,\ \Sigma\ V^T\big) &= \big\{(\Delta X, \Delta Y) \mid \|\Delta X \Sigma\|_F^2 + \|\Delta Y\|_F^2 \le \omega^2\big\}. \end{aligned}$$

Theorem 1 (Minimizer of the ALS variational residual function for matrices) *Let $A \in \mathbb{R}^{n\times m}$ be the current iterate with $r = \operatorname{rank}(A)$ and let $U\Sigma V^T$ be a truncated SVD, $U \in \mathbb{R}^{n\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$, $V^T \in \mathbb{R}^{r\times m}$, of A. Let further C be the variational residual function as in (9). The minimizer X^+ of $\widetilde{X} \mapsto C_{M,P,U\Sigma,V^T}(\widetilde{X}, V^T)$ is given by*

$$X_{i,:}^+ = \underset{\widetilde{X}_{i,:}}{\operatorname{argmin}} \underbrace{\|\widetilde{X}_{i,:} V^T - M_{i,:}\|_{P_{i,:}}^2}_{\text{standard ALS}} + \underbrace{\frac{|P_{i,:}|}{m}\, \omega^2\, \zeta_2\, \|\widetilde{X}_{i,:} \Sigma^{-1}\|_F^2}_{\text{regularization}},$$

where $P_{i,:} := \{p_2^{(k)} \mid p_1^{(k)} = i,\ k = 1, \ldots, |P|\}$ is the corresponding part of the index set P. The minimizer Y^+ of $\widetilde{Y} \mapsto C_{M,P,U,\Sigma V^T}(U, \widetilde{Y})$ is given by

$$Y_{:,j}^+ = \underset{\widetilde{Y}_{:,j}}{\operatorname{argmin}} \underbrace{\|U\ \widetilde{Y}_{:,j} - M_{:,j}\|_{P_{:,j}}^2}_{\text{standard ALS}} + \underbrace{\frac{|P_{:,j}|}{n}\, \omega^2\, \zeta_1\, \|\Sigma^{-1}\ \widetilde{Y}_{:,j}\|_F^2}_{\text{regularization}},$$

where $P_{:,j} := \{p_1^{(k)} \mid p_2^{(k)} = j,\ k = 1, \ldots, |P|\}$. The constants ζ_1 and ζ_2 only depend on the proportions of the representation and sampling set (cf. Remark 2).

The factors $\frac{|P_{i,:}|}{m}$ and $\frac{|P_{:,j}|}{n}$ normalize the penalty terms to the particular magnitudes of the corresponding shares of the sampling set P and hence the standard ALS part. The factors $\zeta_1, \zeta_2 \in (0, 1)$ are in turn independent of i, j, respectively. The ratio of both terms equals the ratio of the mode sizes n, m. The reason for the latter scaling will become apparent in the tensor case (see comments to Theorem 3).

Proof We search for $Y^+ := \operatorname{argmin}_{\widetilde{Y}} C_{M,P,U,\Sigma V^T}(U, \widetilde{Y})$ (the counterpart for X^+ is analogous). Substituting

$$(\Delta X, \Delta Y) \to (\Delta X \Sigma^{-1}, \Delta Y)$$

we can (up to a constant Jacobi determinant) restate C as

$$C_{M,P,U,\Sigma V^T}(U, \widetilde{Y}) \propto \int_{\mathbb{V}_\omega} \|(U + \Delta X \Sigma^{-1})(\widetilde{Y} + \Delta Y) - M\|_P^2 \mathrm{d}\Delta X \mathrm{d}\Delta Y,$$
$$\mathbb{V}_\omega = \{(\Delta X, \Delta Y) \mid \|\Delta X\|^2 + \|\Delta Y\|^2 \le \omega^2\}. \tag{12}$$

Each of the independent columns in the minimizer is restated as

$$Y^+_{:,j} = \operatorname*{argmin}_{\widetilde{Y}_{:,j}} \int_{\mathbb{V}_\omega} \|(U + \Delta X \Sigma^{-1})(\widetilde{Y}_{:,j} + \Delta Y_{:,j}) - M_{:,j}\|^2_{P_{:,j}} \mathrm{d}\Delta X \mathrm{d}\Delta Y. \tag{13}$$

Let j be arbitrary but fixed from now on. For any vector x, it is $\|x\|_{\mathrm{vec}(P_{:,j})} = \|H(j)x\|_F = x^T H(j) x$ for a diagonal, square matrix $H(j) \in \mathbb{R}^{n \times n}$ with $H(j)_{(s),(s)} = \delta_{s \in P_{:,j}}$ (hence $H(j)^2 = H(j)$). Using the normal equation, we obtain $Y^+_{:,j} = W^{-1} b$, where

$$W = \int_{\mathbb{V}_\omega} \left(U + \Delta X \Sigma^{-1}\right)^T H(j) \left(U + \Delta X \Sigma^{-1}\right) \mathrm{d}\Delta X \mathrm{d}\Delta Y$$

and

$$b = \int_{\mathbb{V}_\omega} \left(U + \Delta X \Sigma^{-1}\right)^T H(j) \left(M_{:,j} - \left(U + \Delta X \Sigma^{-1}\right) \Delta Y_{:,j}\right) \mathrm{d}\Delta X \mathrm{d}\Delta Y.$$

In both W and b, any perturbation that appears only one-sided in the expanded product vanishes due to symmetry of $\mathbb{V}_\omega$. Thus $b = |\mathbb{V}_\omega| \, U^T_{P_{:,j},:} M_{P_{:,j},j}$ and

$$W = \int_{\mathbb{V}_\omega} U^T H(j) U \mathrm{d}\Delta X \mathrm{d}\Delta Y + \int_{\mathbb{V}_\omega} (\Delta X \Sigma^{-1})^T H(j) (\Delta X \Sigma^{-1}) \mathrm{d}\Delta X \mathrm{d}\Delta Y$$

Now, let $\ell = \#_X = nr$, $k = \#_Y = rm$. Since $\mathbb{V}$ is a version of the $(\ell + k)$-sphere, we can use the following integration formula: For any $n, m, k \in \mathbb{N}$ let $f : \mathbb{R}^{n+m} \to \mathbb{R}^k$

be a sufficiently smooth function and S_ω^{v-1} be the v-sphere of radius ω. Then

$$\int_{S_\omega^{n+m-1}} f(x_n, x_m)\mathrm{d}x = \int_0^{\pi/2} \omega \int_{S^{n-1}_{\omega\sin(u)}} \int_{S^{m-1}_{\omega\cos(u)}} f(x_n, x_m)\mathrm{d}x_m\mathrm{d}x_n\mathrm{d}u.$$

For a function f we then obtain

$$\begin{aligned}\int_{\mathbb{V}} f\mathrm{d}\Delta X\mathrm{d}\Delta Y &= \int_{\lambda=0}^{\omega}\int_{S_\lambda^{\ell+k-1}} f\mathrm{d}\Delta X\mathrm{d}\Delta Y\mathrm{d}\lambda \\ &= \int_{\lambda=0}^{\omega}\lambda\int_{g=0}^{\pi/2}\int_{S^{k-1}_{\lambda\sin(g)}}\int_{S^{\ell-1}_{\lambda\cos(g)}} f\mathrm{d}\Delta X\mathrm{d}\Delta Y\mathrm{d}g\mathrm{d}\lambda\end{aligned}$$

When f is independent of ΔY, this simplifies to

$$= \int_{\lambda=0}^{\omega}\lambda\int_{g=0}^{\pi/2}|S^{k-1}_{\lambda\sin(g)}|\int_{S^{\ell-1}_{\lambda\cos(g)}} f\mathrm{d}\Delta X\mathrm{d}g\mathrm{d}\lambda$$

We further use the identity

$$\int_0^{\pi/2}\cos(x)^p\sin(x)^q\mathrm{d}x = \frac{\Gamma((p+1)/2)\,\Gamma((q+1)/2)}{2\Gamma((p+q+2)/2)} =: \nu(p,q)$$

We apply these and Lemma 4 for different $f_0 = U^T\,H(j)\,U$ $(\delta=0)$ and $f_1 = (\Delta X\Sigma^{-1})^T\,H(j)\,(\Delta X\Sigma^{-1})$ $(\delta=1)$. For the summands $W(0)+W(1)=W$ this then yields

$$\begin{aligned}W(\delta) &= \int_{\lambda=0}^{\omega}\lambda\int_{g=0}^{\pi/2}\frac{2\pi^{k/2}(\lambda\sin(g))^{k-1}}{\Gamma(k/2)}\left(\lambda^2\cos(g)^2\right)^\delta\frac{2\pi^{\ell/2}(\lambda\cos(g))^{\ell-1}}{\Gamma(\ell/2)}C_H(\delta) \\ &= c\frac{\omega^{k+\ell+2\delta}}{k+\ell+2\delta}\,\nu(\ell-1+2\delta,k-1)C_H(\delta)\end{aligned}$$

for $c = \frac{4\pi^{(k+\ell)/2}}{\Gamma(\ell/2)\Gamma(k/2)}$. The constant matrices C_H are given by

$$\begin{aligned}C_H(0) = \widetilde{C}_H(0) &= U^T_{P_{:,j},:}\,P_{:,j},: \\ |P_{:,j}|^{-1}nrC_H(1) = \widetilde{C}_H(1) &= \Sigma^{-2}\end{aligned}$$

Furthermore, it is $|\mathbb{V}_\omega| = c\frac{\omega^{\ell+k}}{\ell+k}\nu(\ell-1,k-1)$. Factoring out this base volume in $W = |\mathbb{V}_\omega|\widetilde{W}$ by using properties of the Γ function, one derives:

$$\widetilde{W}(0) = \widetilde{C}_H(0),\ \ \widetilde{W}(1) = \frac{|P_{:,j}|}{n}\omega^2\zeta_1\widetilde{C}_H(1),\quad \zeta_1 = \frac{\ell}{r(k+\ell+2)}$$

Restating the result again as a least squares problem finishes the proof. □

Remark 2 *(Specification of constants)* Let $\#_X := nr$, $\#_Y := rm$ be the sizes of the components in the matrix decomposition. The constants in Theorem 1 are given by

$$\zeta_2 = \frac{\#_Y}{r(\#_X + \#_Y + 2)}, \quad \zeta_1 = \frac{\#_X}{r(\#_X + \#_Y + 2)}.$$

However, when changing the rank, this would impose a slight offset in continuity of both $f_{\mathscr{M}^{(1)*}}$ and $f_{\mathscr{M}^{(2)*}}$. This problem is simply resolved by substituting ω by $\widetilde{\omega}$ properly for each value r, such that we can set and normalize

$$\omega^2\zeta_2 = \widetilde{\omega}^2\frac{m}{n+m}, \quad \omega^2\zeta_1 = \widetilde{\omega}^2\frac{n}{n+m}.$$

We will still just write ω although we replace ζ_1 and ζ_2 by the adapted values.

Manipulating ω does not change the representation independency of the two methods $(\mathscr{M}^{(1)})^*$ and $(\mathscr{M}^{(2)})^*$, since for fixed r, the value $\omega > 0$ just remains an arbitrary constant (cf. Lemma 7). We arrive at the main result for the matrix case:

Theorem 2 (Stability of the variational matrix methods) *The methods* $(\mathscr{M}^{(1)})^*$ (10) *and* $(\mathscr{M}^{(2)})^*$ (11) *are stable.*

Proof Follows as special case from the proof of the tensor version, Theorem 4. □

Despite the technicalities involved in the proof, the simplicity of the idea becomes apparent by setting $P = \mathscr{I}$ (being analogous to the argumentation in Sect. 2.1). For a certain constant $c \in \mathbb{R}$, in the setting of Theorem 1, we then have

$$\begin{aligned} f_{\mathscr{M}^{(2)*}}(U\Sigma V^T) &= U \cdot \underbrace{(I + c\Sigma^{-2})^{-1}}_{\text{regularization}} \cdot \underbrace{U^T M}_{\text{standard ALS}} \\ &= \left(\left(1 + c\sigma_1^{-2}\right)^{-1} \cdot U_{:,1}U_{:,1}^T M, \ \ldots, \ \left(1 + c\sigma_r^{-2}\right)^{-1} \cdot U_{:,r}U_{:,r}^T M \right). \end{aligned} \tag{14}$$

If now $\sigma_r \to 0$, then also $(1 + c\sigma_r^{-2})^{-1} \to 0$ and we obtain the same result as if we would have truncated the representation $(X, Y) = (U, \Sigma V^T)$ to rank $r - 1$ beforehand. We also denote these additional factors as *filter*, as they filter out influence corresponding to low singular values. Note that obtaining small singular values is not penalized, but using components corresponding to small ones is.

Algorithm 1 performs one stable ALS approximation (SALSA) sweep, that is it applies the two stable methods from Theorem 2. Before each update, any current singular value $\sigma_i < \sigma_{\min}$ is replaced by $\sigma_{\min}$, which in turn is set as fraction $f_{\min} \ll 1$ of the current residual (Algorithm 2). Although the influence of this manipulation on the subsequent step is thereby marginal, it is necessary since otherwise σ_i may irreversibly converge to zero (cf. Remark 1, for more details, see Sect. 6).

Algorithm 1 Stable Matrix Completion

Require: limit $\sigma_{\min}$, parameter ω, initial guess $A = XY \in \mathbb{R}^{n\times m}$ such that X contains the left singular vectors of A, and data points $M|_P$

1: compute the SVD $U\Sigma V^T := Y$ and update $\sigma_i := \max(\sigma_i, \sigma_{\min}),\ i = 1, \dots, r$
2: set $X := XU\Sigma$ and $Y := V^T$
3: for $i = 1, \dots, n$ update

$$X_{i,:} := \underset{\widetilde{X}_{i,:}}{\operatorname{argmin}} \|\widetilde{X}_{i,:}Y - M_{i,:}\|^2_{P_{i,:}} + \frac{|P_{i,:}|}{m}\frac{\omega^2 m}{n+m}\|\widetilde{X}_{i,:}\Sigma^{-1}\|_F^2 \tag{15}$$

4: compute the SVD $U\Sigma V^T := X$ and update $\sigma_i := \max(\sigma_i, \sigma_{\min}),\ i = 1, \dots, r$
5: set $X := U$ and $Y := \Sigma V^T Y$
6: for $j = 1, \dots, m$ update

$$Y_{:,j} := \underset{\widetilde{Y}_{:,j}}{\operatorname{argmin}} \|X\widetilde{Y}_{:,j} - M_{:,j}\|^2_{P_{:,j}} + \frac{|P_{:,j}|}{n}\frac{\omega^2 n}{n+m}\|\Sigma^{-1}\widetilde{Y}_{:,j}\|_F^2 \tag{16}$$

3.2 Rank adaption

The key aspect of stability is that it renders explicit rank adaption near unnecessary, since the optimization relies on the magnitude of singular values, as (14) suggests. Starting from an initial representation and a large value ω proportional to the norm of the iterate A, the parameter ω is decreased after each iteration. The singular values are devided into two types:

Definition 4 (*Stabilized rank and minor singular values*) A singular value σ_i is called *stabilized*, if it is larger than a certain fixed fraction of ω [meaning any corresponding terms have an increased influence, cf. (14)]. Otherwise, it is called minor (as a removal of such does not notably change the next steps). The *stabilized rank* only counts the number of stabilized singular values.

The actual rank is only modified as to make sure that there is always a fixed, small amount of minor singular values, i.e.

$$|\{i \mid 0 < \sigma_i < f_{\text{minor}} \cdot \omega\}| \overset{!}{=} k_{\text{minor}}, \tag{17}$$

for constants $f_{\text{minor}} < 1$ and $k_{\text{minor}} \in \mathbb{N}$. This states the basic concept of implicit rank adaption and we will provide a more detailed discussion in the later Sect. 8 for the elaborate tensor case. For performance evaluation, a validation set may be used:

Definition 5 (*Validation set*) For a given P, the sampling or training set, we define $P_2 \subset P$ as validation set. This set may be chosen randomly or specifically distributed. The actual set used for the optimization is replaced by $P \leftarrow P \setminus P_2$ (keeping the same symbol).

The stopping criteria in Algorithm 2 may depend on the behavior of P_2, or may simply be based on a rank bound, e.g. $r \le |P|/(n+m)$. The latter criterion, however, only suffices in the matrix case.

Algorithm 2 SALSA Algorithm

Require: $P \subset \mathscr{I}$, $M|_P$
1: initialize X, Y s.t. $\tau_r(X,Y) \equiv \|M|_P\|_1/|P|$ for $r \equiv 1$ and $\omega = 1/2 \cdot \|\tau_r(X,Y)\|_F$
2: split off a small validation set $P_2 \subset P$ for performance evaluation
3: **for** iter $= 1, 2, \ldots$ **do**
4: proceed SALSA sweep* (Algorithm 1)
5: *: and renew lower limit $\sigma_{\min} := f_{\sigma_{\min}} \cdot \frac{|\mathscr{I}|}{|P|}\|\tau_r(X,Y) - M\|_P$
6: *: and decrease ω
7: adapt rank according to (17) (start this when the first few iteration have passed)
8: **if** a stopping criterion applies **then**
9: terminate algorithm
10: **return** iterate for which $\|\tau_r(X,Y) - M\|_{P_2}$ was lowest
11: **end if**
12: **end for**

Remark 3 *(On convergence estimates)* Due to the specific dependency of the regularization on the current singular values of the iterate, it may be impossible to restate the problem as minimization of a modified, global cost function. Furthermore, the iterate does not remain on a fixed manifold of low rank. Given that also rank deficient points may be included in the optimization due to the stability, theoretical convergence results so far remain subject to future work.

4 Generalization to high-dimensional tensors

For the rest of this article we consider the problem of (approximately) reconstructing a *tensor* $M \in \mathbb{R}^{\mathscr{I}}$ from a given data set $M|_P = \{M_p\}_{p\in P}$, where now

$$P \subset \mathscr{I} := \bigtimes_{\mu=1}^{d} \mathscr{I}_\mu, \quad \mathscr{I}_\mu := \{1, \ldots, n_\mu\},\ \mu = 1, \ldots, d. \tag{18}$$

We further generalize the representation map and data space $\tau_r : \mathscr{D}_r \to \mathscr{D}_r$ as well as the rank r together with the spaces $\mathscr{T}_r$ to the tensor train (TT-)format (Definition 7, [33,34], also called Matrix Product States (MPS) [41,45] or interpreted as a special case of the Hierarchical Tucker format [14,19]). Stability for tensor methods is thereby defined through Definition 1 as well. For the underlying tensor M it is now assumed that its TT-singular values (cf. Definition 6) decline sufficiently fast in order to yield an approximation $M_\varepsilon \in \mathscr{T}_r$.

Definition 6 (*TT-singular values and TT-rank*) Analogously to the TT-ranks, we define the TT-singular values $\sigma = (\sigma^{(1)}, \ldots, \sigma^{(d-1)})$ of a tensor A as the unique singular values of the corresponding matricizations $A^{(\{1,\ldots,\mu\})} \in \mathbb{R}^{(\mathscr{I}_1\times\cdots\times\mathscr{I}_\mu)\times(\mathscr{I}_{\mu+1}\times\cdots\times\mathscr{I}_d)}$ with entries

$$A^{(\{1,\ldots,\mu\})}((i_1, \ldots, i_\mu),\ (i_{\mu+1}, \ldots, i_d)) := A(i)$$

of A, such that $\sigma^{(\mu)}$ contains the ordered singular values of $A^{(\{1,\dots,\mu\})}$, $\mu = 1, \dots, d-1$. The TT-rank r_μ is the number of nonzero TT-singular values in $\sigma^{(\mu)}$. We also call $\sigma^{(\mu)}$ the μ-th singular values.

Definition 7 (*TT-format*) A tensor $A \in \mathbb{R}^{\mathscr{I}}$ is in the set $\mathscr{T}_{\leq r}$, often also denoted $\mathrm{TT}(\leq r)$, $r \in \mathbb{N}^{d-1}$, if for $\mu = 1, \dots, d$ and $i_\mu \in \mathscr{I}_\mu$ there exist $G_\mu(i_\mu) \in \mathbb{R}^{r_{\mu-1} \times r_\mu}$ ($r_0 = r_d = 1$) such that

$$A = \tau_r(G), \quad A(i_1, \dots, i_d) := G_1(i_1) \cdots G_d(i_d), \qquad i \in \mathscr{I}.$$

The representation of A in this form is shortly called the TT format. The single G_μ, as well as similarly structured objects, are called cores (or sometimes nodes).

The TT-SVD [33] provides that if r is the TT-rank of A, then it follows $A \in \mathscr{T}_r$. So G is also called low-rank representation of A. The function τ_r will appear in different forms, each denoting the mapping from a type of representation of rank r to the represented object.

Definition 8 (*Unfoldings*) For a core H (possibly a product of smaller cores in the TT representation) with $H(i) \in \mathbb{R}^{k_1 \times k_2}$, $i = 1, \dots, n$, we denote the left unfolding $\mathfrak{L}(H) \in \mathbb{R}^{k_1 \cdot n \times k_2}$, in which the matrices contained in the core are stacked below each other, and right unfolding $\mathfrak{R}(H) \in \mathbb{R}^{k_1 \times k_2 \cdot n}$, in which they are stacked side by side, by

$$(\mathfrak{L}(H))_{(\ell,j),q} := (H(j))_{\ell,q}\,, \quad (\mathfrak{R}(H))_{\ell,(q,j)} := (H(j))_{\ell,q}\,,$$

for $1 \leq j \leq n$, $1 \leq \ell \leq k_1$ and $1 \leq q \leq k_2$.

Our goal is to determine the ranks adaptively. We will demonstrate why this can be even more troublesome than in the matrix case in the following.

For many tensors that stem from practical applications one observes exponentially decaying singular values, yet the rate of decay may strongly vary for the $d-1$ different matricizations. Theoretically, there is no non-trivial limitation to the shape of the TT-singular values as described in following lemma.

Lemma 5 (Feasibility of TT-singular values [26]) *Let $\sigma = (\sigma^{(1)}, \dots, \sigma^{(d-1)})$ be a $(d-1)$-tuple of weakly decreasing r_μ-tuples, $\mu = 1, \dots, d-1$, for which $\|\sigma^{(i)}\|_2 = \|\sigma^{(j)}\|_2$ for all $i, j = 1, \dots, d-1$. Then there exist mode sizes $n_1, \dots, n_d \in \mathbb{N}$ and a tensor $A \in \mathbb{R}^{\mathscr{I}}$ such that this tensor has TT-singular values σ.*

The following exemplary tensor emphasizes the problems heuristics encounter in the rank adaption of a tensor with inconvenient singular values.

Example 3 *(Rank adaption test tensor)* For $k \in \mathbb{N}$, let $Q \in \mathbb{R}^{n_1 \times \dots \times n_4}$ be an orthogonally decomposable 4-dimensional Tensor with TT-rank (k, k, k) and uniform singular values $\sigma^{(1)} = \sigma^{(2)} = \sigma^{(3)} = (\alpha, \alpha, \dots)$ as well as $B \in \mathbb{R}^{n_5 \times n_6}$ be a rank $2k$ matrix with exponentially decaying singular values $\sigma^{(5)} \propto (\beta^{-1}, \beta^{-2}, \dots)$ for some $\alpha, \beta > 0$. Then the separable tensor $M \in \mathbb{R}^{n_1 \times \dots \times n_6}$ defined by $M(i) = Q(i_1, \dots, i_4) \cdot B(i_5, i_6)$ has singular values σ and rank $r^{(M)} = (k, k, k, 1, 2k)$.

By its definition, M is separable into a 4- and a 2-dimensional tensor (Q, B). Knowing this would of course drastically simplify the problem. We now consider the performance of two very basic rank adaption ideas.

1. *Greedy, single rank increase*: We test for maximal improvement by increase of one of the ranks r_μ $(\mu = 1, \dots, d-1)$ of the iterate starting from $r \equiv 1$. Solely increasing either of r_2, r_3 or r_4 will give close to no improvement. As further shown in [12], the approximation of orthogonally decomposable tensors with lower rank can be problematic. In numerical tests (cf. Sect. 9.8), we can observe that r_5 is often increased to a maximum first. Thereby, extremely small singular values are involved that lie far beneath the current approximation error, although the rank is not actually overestimated.
2. *Uniform rank increase and coarsening*: We increase every rank r_μ $(\mu = 1, \dots, d-1)$ starting from $r \equiv 1$ and decrease ranks when the corresponding singular values are below a threshold. The problem with this strategy is quite plain, namely that for the target tensor M, it holds $r_5^{(M)} = 1$. If this rank is overestimated, the observed sampling points will be misinterpreted (overfitting) and it does not matter how small corresponding singular values become (see Lemma 2).

These indicated difficulties gain more importance with high dimension, but for one microstep at a time, can be resolved by regarding only three components of a tensor. We describe this in the following Sect. 4.1.

The tensor from Example 3 can also be constructed explicitly. We define a representation G for $A = \tau_r(G)$ via left and right unfoldings by

$$\begin{aligned} &\mathfrak{L}(G_1) := Q_1, \\ &G_2(i_2) = G_3(i_3) := I_k, \qquad 1 \le i_2 \le n_2, \quad 1 \le i_3 \le n_3, \\ &\mathfrak{R}(G_4) := Q_4^T, \ \mathfrak{L}(G_5) := Q_5, \ \mathfrak{R}(G_6) := \Sigma_5 Q_6^T, \end{aligned}$$

for (column-) orthogonal matrices $Q_1 \in \mathbb{R}^{n_1 \times r_1}$, $Q_4 \in \mathbb{R}^{n_4 r_4 \times r_3}$, $Q_5 \in \mathbb{R}^{r_4 n_5 \times r_5}$ $Q_6 \in \mathbb{R}^{n_6 \times r_5}$ and $(\sigma_5)_i \propto \beta^{-i}$, $\beta > 1$. This tensor has exactly the properties postulated in the example.

4.1 Notations and reduction to three dimensions

For necessary simplicity, we reduce the d dimensional setting to a three dimensional one:

Definition 9 (*Interface matrices*) Let $\mu \in \{1, \dots, d\}$. For a representation G, we define the left interface matrix $G^{<\mu} \in \mathbb{R}^{n_1 \dots n_{\mu-1} \times r_{\mu-1}}$ (cf. Definitions 7) via

$$G^{<\mu}_{(i_1,\dots,i_{\mu-1}),:} := G_1(i_1) \dots G_{\mu-1}(i_{\mu-1}),$$

as well as the right interface matrix $G^{>\mu} \in \mathbb{R}^{r_\mu \times n_{\mu+1} \dots n_d}$ via

$$G^{>\mu}_{:,(i_1,\dots,i_{\mu-1})} := G_1(i_{\mu+1}) \dots G_d(i_d).$$

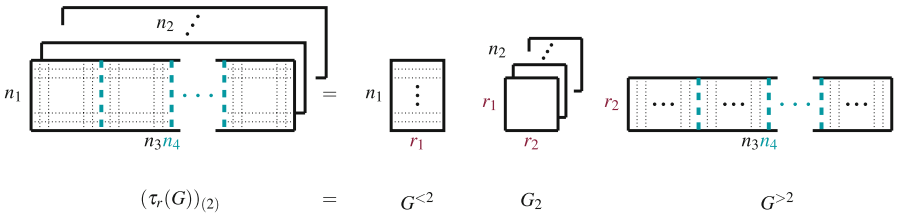

Fig. 3 The decomposition of a core unfolding with respect to 2 of a four dimensional tensor into the left and right interface matrices as well as the intermediate core

We further define the core $A_{(\mu)} \in \left(\mathbb{R}^{n_1 \dots n_{\mu-1} \times n_{\mu+1} \dots n_d}\right)^{\mathscr{I}_\mu}$ as core unfolding with respect to mode μ of a tensor A by

$$A_{(\mu)}(i_\mu)_{(i_1,\dots,i_{\mu-1}),(i_{\mu+1},\dots,i_d)} = A(i). \tag{19}$$

For any representation it hence holds $(\tau_r(G))_{(\mu)}(j) = G^{<\mu}\ G_\mu(j)\ G^{>\mu}$, $j = 1, \dots, n_\mu$. Multiplication of a core H with a matrix B yields again cores, HB and BH, given by

$$(HB)(j) := H(j)B \quad \text{and} \quad (BH)(j) := BH(j),$$

respectively, for all possible j. The previous notations then allow us to compactly write

$$(\tau_r(G))_{(\mu)} = G^{<\mu}\ G_\mu\ G^{>\mu}. \tag{20}$$

This relation is displayed in Fig. 3.

The interface matrices equal left and right unfoldings, respectively:

$$\begin{aligned} G^{<\mu} &= \mathfrak{L}(G_{1,\dots,\mu-1}), \quad G_{1,\dots,\mu-1}((i_1,\dots,i_{\mu-1})) := G_1(i_1)\dots G_{\mu-1}(i_{\mu-1}), \\ G^{>\mu} &= \mathfrak{R}(G_{\mu+1,\dots,d}), \quad G_{\mu+1,\dots,d}((i_{\mu+1},\dots,i_d)) := G_1(i_{\mu+1})\dots G_d(i_d). \end{aligned} \tag{21}$$

In terms of Definition 3, for $P = \{p^{(i)} \mid i = 1, \dots, |P|\}$ and the operation $(\cdot)_{(\mu)}$ (19), which is used to combine components $s = 1, \dots, \mu-1$ as well as $s = \mu+1, \dots, d$, let

$$P_{(\mu)} = \left\{ \left(\left(p_1^{(i)}, \dots, p_{\mu-1}^{(i)}\right), p_\mu^{(i)}, \left(p_{\mu+1}^{(i)}, \dots, p_d^{(i)}\right) \right) \mid i = 1, \dots, |P| \right\}.$$

Thereby, $A|_P$ contains the same entries as $(A_{(\mu)})|_{P_{(\mu)}}$. For the selection of one slice, $(\cdot)_{(\mu)}(j)$, we denote

$$P_{(\mu)}(j) = \left\{ \left(\left(p_1^{(i)}, \dots, p_{\mu-1}^{(i)}\right), \left(p_{\mu+1}^{(i)}, \dots, p_d^{(i)}\right) \right) \mid p_\mu^{(i)} = j,\ i = 1, \dots, |P| \right\}. \tag{22}$$

Likewise, the vectorization of an index set $S \subset \mathbb{R}^{n\times m}$ is defined by $\mathrm{vec}(S) = \{s_1 + n(s_2 - 1) \in \mathbb{R} \mid s \in \mathbb{R}^2\}$.

Without loss of generality we can restrict our consideration to three dimensional tensors that correspond to the left and right interface matrices as well as the respective intermediate cores (cf. Definition 8):

Remark 4 *(Reduction to three dimensions)* When $\mu \in \{1, \dots, d\}$ is fixed, we will only use the short notations

- $(L,\ N,\ R) = (G^{<\mu},\ G_\mu,\ G^{>\mu})$
- $(n_L,\ n_N,\ n_R) = (n_1 \cdot \ldots \cdot n_{\mu-1},\ n_\mu,\ n_{\mu+1} \cdot \ldots \cdot n_d)$
- $(\gamma,\ \theta) = (\sigma^{(\mu-1)},\ \sigma^{(\mu)})$ and $(\Gamma,\ \Theta) = (\Sigma^{(\mu-1)},\ \Sigma^{(\mu)}) = (\mathrm{diag}(\sigma^{(\mu-1)}), \mathrm{diag}(\sigma^{(\mu)}))$
- $(r_\gamma,\ r_\theta) = (r_{\mu-1},\ r_\mu)$
- $B = M_{(\mu)}$ and $S = P_{(\mu)}$

Hence, the important variables are L (left part), R (right part), N (new part), B (right hand side) and S (sampling).

The microsteps $\mathscr{M}^{(1)}, \dots, \mathscr{M}^{(d)}$ of ALS for the tensor train format only change the respective G_μ and are given by

$$\begin{aligned} \mathscr{M}_r^{(\mu)}(G) &:= (G_1, \dots, G_{\mu-1}, G_\mu^+, G_{\mu+1}, \dots, G_d) \\ G_\mu^+ &:= \operatorname*{argmin}_{G_\mu} \|\tau_r(G) - M\|_P = \operatorname*{argmin}_{\widetilde{N}} \|L \cdot \widetilde{N} \cdot R - B\|_S \end{aligned} \tag{23}$$

or equivalently $G_\mu^+(j) = \operatorname{argmin}_{\widetilde{N}(j)} \|L \cdot \widetilde{N}(j) \cdot R - B(j)\|_{S(j)}$—an equation in which only matrices are involved. We only need to consider three-dimensional tensors $A \in \mathbb{R}^{n_L \times n_N \times n_R}$, with TT-rank (r_γ, r_θ) and TT-singular values (γ, θ). For simplicity, we redefine τ_r for this case via $A = \tau_r(L, N, R)$.

5 Stable alternating least squares microsteps for tensor completion

With the derivation in Sect. 4.1, we have seen that it is sufficient to only consider ALS for order 3 tensors.

Definition 10 (*Variational residual function (cf. Definition 2)*) Let $\omega \geq 0,\ s_1, s_2 > 0$ and B, S, L, N, R as in Definition 4. We define the variational residual function $C := C_{B,S,L,N,R}$ for $\mathbb{V}_\omega := \mathbb{V}_\omega(L, N, R)$ by

$$\begin{aligned} C(\widetilde{N}) &:= \int_{\mathbb{V}_\omega(L,N,R)} \|(L + s_1 \Delta L)(\widetilde{N} + \Delta N)(R + s_2 \Delta R) - B\|_S^2 \mathrm{d}\Delta L \mathrm{d}\Delta N \mathrm{d}\Delta R, \\ \mathbb{V}_\omega &:= \{(\Delta L, \Delta N, \Delta R) \mid \|\Delta L N R\|_F^2 + \|L \Delta N R\|_F^2 + \|L N \Delta R\|_F^2 \leq \omega^2\}, \end{aligned} \tag{24}$$

where s_1, s_2 are scalings that only depend on the proportions of the representation, to be specified by Lemma 12.

As in the matrix case, ΔN does not influence the minimizer, so we omit it from now on. It should further be noted that $\mathbb{V}_\omega$ does **not** depend on the unknown $\widetilde{N}$. The scalings s_1, s_2 will become relevant in Sect. 7 in order to achieve a similar effect as in Remark 2.

5.1 Standard representation of a TT-tensor

A representation $G = (L, N, R)$ can be changed without changing the generated tensor $A = \tau_r(G)$ [23,37], more specifically

$$\tau_r(G) = \tau_r(\widetilde{G}) \quad \Leftrightarrow \quad \widetilde{G} = (\widetilde{L}, \widetilde{N}, \widetilde{R}) = \left(LT_1^{-1}, T_1 N T_2^{-1}, T_2 R\right) \tag{25}$$

for two regular matrices $T_1 \in \mathbb{R}^{r_\gamma \times r_\gamma}$, $T_2 \in \mathbb{R}^{r_\theta \times r_\theta}$. One can define an extended standard representation that explicitly contains the unique TT-singular values (γ, θ) which is essentially unique (in terms of uniqueness of the truncated matrix SVD[2]). For the construction, a slightly modified TT-SVD [33] is used.[3] An analogous decomposition also appeared earlier in [41] and is, within the quantum computing community, sometimes referred to as *canonical MPS form* (not to be confused with canonical polyadic decomposition). This normalization is needed to obtain the same simplification as in Theorem 1.

Lemma 6 (Standard representation) *Let $A \in \mathbb{R}^{n_L \times n_N \times n_R}$ be a tensor. There exists an essentially unique (extended) representation*

$$\mathscr{G} = (\mathscr{L}, \Gamma, \mathscr{N}, \Theta, \mathscr{R}) \tag{26}$$

for which $A = \tau_r(\mathscr{L},\ \Gamma\mathscr{N}\Theta,\ \mathscr{R})$ as well as $\mathscr{L}\ \Gamma\ \mathfrak{R}(\mathscr{N}\Theta\mathscr{R})$ and $\mathfrak{L}(\mathscr{L}\Gamma\mathscr{N})\ \Theta\ \mathscr{R}$ are (truncated) SVDs of $A^{(\{1\})}$ and $A^{(\{1,2\})}$, respectively. This in turn implies that $\mathscr{L}$ and $\mathfrak{L}(\Gamma\mathscr{N})$ are column orthogonal, as well as $\mathscr{R}$ and $\mathfrak{R}(\mathscr{N}\Theta)$ are row orthogonal.

Proof 1. *Uniqueness:* Let there be two such representations $\widetilde{\mathscr{G}}$ and $\mathscr{G}$. Since the left-singular vectors of $A^{(\{1\})}$ are essentially unique, we conclude $\widetilde{\mathscr{L}} = \mathscr{L}W_1$ for an orthogonal matrix W_1 that commutes with Γ. Via an SVD of $A^{(\{1,2\})}$ it follows that $\widetilde{\mathscr{R}} = W_2^T\mathscr{R}$ for an orthogonal matrix W_2 that commutes with Θ. Furthermore $\mathfrak{L}(\mathscr{L}\Gamma W_1\widetilde{\mathscr{N}}) = \mathfrak{L}(\widetilde{\mathscr{L}}\Gamma\widetilde{\mathscr{N}}) = \mathfrak{L}(\mathscr{L}\Gamma\mathscr{N})W_2$. The map $x \mapsto \mathfrak{L}(\mathscr{L}\Gamma x)$ is linear and, in this case, of full rank. This implies $\widetilde{\mathscr{N}} = W_1^T\mathscr{N}W_2$.

2. *Existence (constructive):* Let $A = \tau_r(\widetilde{L}, \widetilde{N}, \widetilde{R})$ where $\mathfrak{R}(\widetilde{N})$ and $\widetilde{R}$ are column orthogonal [this can always be achieved using (25)]. An SVD of $\widetilde{L}$ yields $\widetilde{L} = \mathscr{L}\ \Gamma\ V_1^T$, since $\mathscr{L}\ \Gamma\ \mathfrak{R}(V_1^T\widetilde{N}\widetilde{R})$ is a truncated SVD of $A^{(\{1\})}$. A subsequent SVD of $\mathfrak{L}(\Gamma V_1^T\widetilde{N})$ yields $\Gamma V_1^T\widetilde{N} = \widehat{N}\ \Theta\ V_2^T$, since $\mathfrak{L}(\mathscr{L}\widehat{N})\ \Theta\ (V_2^T\widetilde{R})$ is a truncated SVD of $A^{(\{1,2\})}$. We can finish the proof defining $\mathscr{N} := \Gamma^{-1}\widehat{N}$ and $\mathscr{R} = V_2^T\widetilde{R}$. Note that, by construction, $\mathfrak{L}(\Gamma\mathscr{N})$ is column-orthogonal.

[2] Both $U\Sigma V^T$ and $\widetilde{U}\Sigma\widetilde{V}^T$ are truncated SVDs of A if and only if there exists an orthogonal matrix W that commutes with Σ and for which $\widetilde{U} = UW$ and $\widetilde{V} = VW$. For any subset of pairwise distinct nonzero singular values, the corresponding submatrix of W needs to be diagonal with entries in $\{-1, 1\}$.

[3] Although called TT-SVD, unlike the matrix SVD, decompositions constructed by the algorithm do not explicitly contain the singular values.

3. *Implied orthogonality:* Using the essential uniqueness, it follows that $\mathfrak{L}(\Gamma\mathscr{N})$ must indeed be column-orthogonal. By analogously constructing the extended representation from right to left we would obtain that $\mathfrak{R}(\mathscr{N}\Theta)$ is row-orthogonal. By uniqueness it follows again that this is always the case. □

Remark 5 *(Conventional form of standard representation)* Throughout the rest of the article, the standard representation will mostly appear in form of a specific, conventional representation

$$(L,\ N,\ R) = (\mathscr{L},\ \Gamma\mathscr{N}\Theta,\ \mathscr{R}), \tag{27}$$

hence with interface matrices $\mathscr{L}$ and $\mathscr{R}$ given by corresponding singular vectors.

5.2 Minimizer of the variational residual function

We define (from now on) our method as

$$\mathscr{M}^*(L,\ N,\ R) = \left(L,\ \operatorname*{argmin}_{\widetilde{N}} C(\widetilde{N}),\ R\right) \tag{28}$$

with $C = C_{B,S,L,N,R}$ as in (24). Although Theorem 3, or more specifically the regularity of $Y(j)$ given by (30), later provide the uniqueness of the minimizer, we up to that point formally use the minimization of $\|\tau_r(L, \widetilde{N}, R)\|_F$ as secondary and representation independent criterion. The special cases $\mu \in \{1, d\}$ can be derived from the general case (Remark 8) and comply with the matrix case.

Lemma 7 (Representation independent) *The method $\mathscr{M}^*$ is representation independent.*

Proof Let $N^+ := \operatorname{argmin}_{\widetilde{N}} C$, $C = C_{B,S,L,N,R}(\widetilde{N})$ and $\widehat{N}^+ := \operatorname{argmin}_{\widetilde{N}} \widehat{C}$, $\widehat{C} = C_{B,S,\widehat{L},\widehat{N},\widehat{R}}(\widetilde{N})$ for representations $\tau_r(L, N, R) = \tau_r(\widehat{L}, \widehat{N}, \widehat{R})$ as well as $\widehat{\mathbb{V}}_\omega = \mathbb{V}_\omega(\widehat{L}, \widehat{N}, \widehat{R})$ and $\mathbb{V}_\omega = \mathbb{V}_\omega(L, N, R)$. According to (25), there exist two matrices T_1, T_2 such that

$$\left(LT_1, T_1^{-1}NT_2, T_2^{-1}R\right) = (\widehat{L}, \widehat{N}, \widehat{R}).$$

Hence

$$\widehat{C}(\widetilde{N}) = \int_{\widehat{\mathbb{V}}_\omega} \left\| \left(L + s_1 \Delta\widehat{L} T_1^{-1}\right) T_1 \widetilde{N} T_2^{-1} \left(R + s_2 T_2 \Delta\widehat{R}\right) - B \right\|_S^2 \mathrm{d}\Delta\widehat{L}\mathrm{d}\Delta\widehat{N}\mathrm{d}\Delta\widehat{R},$$

$$\text{with } \widehat{\mathbb{V}}_\omega = \left\{ (\Delta\widehat{L}, \Delta\widehat{N}, \Delta\widehat{R}) \mid \|\Delta\widehat{L}T_1^{-1}NR\|_F^2 + \|LT_1^{-1}\Delta\widehat{N}T_2R\|_F^2 + \|LNT_2\Delta\widehat{R}\|_F^2 \le \omega^2 \right\}$$

The substitution $(\Delta\widehat{L}, \Delta\widehat{N}, \Delta\widehat{R}) \overset{\iota}{\to} (\Delta L T_1, T_1^{-1}\Delta N T_2, T_2^{-1}\Delta R)$ introduces a constant Jacobi Determinant $|\det(J_\iota)| = 1$. We obtain

$$\begin{aligned}\widehat{C}(\widetilde{N}) &:= \int_{\mathbb{V}_\omega} \left\| \left(L + s_1\Delta L\right)\left(T_1\widetilde{N}T_2^{-1}\right)\left(R + s_2\Delta R\right) - B \right\|_S^2 \mathrm{d}\Delta L \mathrm{d}\Delta N \mathrm{d}\Delta R \\ &= C\left(T_1\widetilde{N}T_2^{-1}\right)\end{aligned}$$

Hence $\widehat{N}^+ = T_1^{-1}N^+T_2$. This is the same relation given for N and $\widehat{N}$ and therefore $\tau_r(L, N^+, R) = \tau_r(\widehat{L}, \widehat{N}^+, \widehat{R})$ (which is a set equality if the minimizer is not assumed to be unique). □

Corollary 1 (Integral over Kronecker product) *Let $\omega_1 > 0$.*
Further, let $H \in \mathbb{R}^{(n_X n_Y)\times(n_X n_Y)}$ as well as $Y \in \mathbb{R}^{n_Y \times n_Y}$ be matrices and

$$V_{\omega_1}^{(n_X, m_X)} = \{X \in \mathbb{R}^{n_X \times m_X} \mid \|X\|_F = \omega_1\}.$$

Then

$$\int_{V_{\omega_1}^{(n_X,m_X)}} (X \otimes Y)^T H (X \otimes Y) \mathrm{d}X = \frac{\omega_1^2 |V_{\omega_1}^{(n_X,m_X)}|}{n_X m_X} I_{m_X} \otimes Y^T H^* Y$$

for $H^* = \mathrm{tr}_1(H) \in \mathbb{R}^{n_Y \times n_Y}$.[4] *For an analog $V_{\omega_2}^{(n_Y,m_Y)}$, $\omega_2 > 0$, we further have*

$$\iint_{V_{\omega_1}^{(n_X,m_X)}, V_{\omega_2}^{(n_Y,m_Y)}} (X \otimes Y)^T H (X \otimes Y) \mathrm{d}X \mathrm{d}Y = \frac{\omega_1^2\omega_2^2 |V_{\omega_1}^{(n_X,m_X)}||V_{\omega_2}^{(n_Y,m_Y)}|}{n_X m_X n_Y m_Y} tr(H) I_{m_X m_Y}.$$

Proof Using the splitting $H = \sum_{i,j} h_{i,j} \otimes e_i e_j^T$, $h_{i,j} \in \mathbb{R}^{n_X \times n_X}$, Lemma 4 can be applied to each summand, separately for X and Y. □

We now derive the minimizer of the variational residual function (24). Due to Lemma 7, we can use the standard representation in form of Remark 5 for simplification. In this case, $\mathbb{V}_\omega$ takes the convenient form

$$\begin{aligned}&\mathbb{V}_\omega(\mathscr{L},\ \Gamma\ \mathscr{N}\ \Theta,\ \mathscr{R}) \\ &\quad = \left\{(\Delta L, \Delta N, \Delta R) \mid \|\Delta L \Gamma\|_F^2 + \|\Delta N\|_F^2 + \|\Theta \Delta R\|_F^2 \le \omega^2\right\}. \end{aligned} \tag{29}$$

Theorem 3 (Minimizer of the ALS variational residual function) *Let $(\mathscr{L},\ \Gamma,\ \mathscr{N},\ \Theta,\ \mathscr{R})$ be the standard representation (26) for a tensor A. The minimizer N^+ of the residual function $C_{B,S,\mathscr{L},\Gamma\mathscr{N}\Theta,R}$ as in (24) is given by*

$$N^+(j) = \underset{\widetilde{N}(j)}{\mathrm{argmin}} \underbrace{\| \mathscr{L}\, \widetilde{N}(j)\, \mathscr{R} - B(j) \|_{S(j)}^2}_{\text{standard ALS}} + \underbrace{\Omega(j),}_{\text{(regularization)}} \quad j = 1, \ldots, n_N$$

[4] tr_1 is the partial trace, i.e. $(H^*)_{i,j} = \mathrm{tr}(h_{i,j})$, $H = \sum_{i,j} h_{i,j} \otimes e_i e_j^T$, $h_{i,j} \in \mathbb{R}^{n_X \times n_X}$.

$$\Omega(j) = \begin{cases} & n_L^{-1}\zeta_1\ \omega^2 \| \ \Gamma^{-1} \quad \widetilde{N}(j)\ \mathscr{R}_{:,S(j)_2}\ \|_F^2 \\ + & n_R^{-1}\zeta_2\ \omega^2 \| \ \mathscr{L}_{S(j)_1,:}\ \widetilde{N}(j) \quad \Theta^{-1}\ \|_F^2 \\ + & |S(j)|(n_R n_L)^{-1}\zeta_{(1,2)}\ \omega^4 \| \ \Gamma^{-1} \quad \widetilde{N}(j) \quad \Theta^{-1}\ \|_F^2 \end{cases} \tag{30}$$

with $S(j)_u = (x_u^{(1)}, x_u^{(2)}, \ldots)$, $u = 1, 2$, *for* $S(j) = \{x^{(1)}, x^{(2)}, \ldots, x^{(|S(j)|)}\} \subset \mathbb{N}^2$. *The constants* ζ *only depend on the proportions of the representation and sampling set (cf. Remark 6) as well as the constant scalings* s_1, s_2.

The regularization term is more straightforward than it might appear. The computational complexity for the calculation of the minimizer is of the same order (with near same constant) as for standard ALS, for which the matrices $\mathscr{L}_{S(j)_1,:} \in \mathbb{R}^{|S(j)|\times r_\gamma}$ and $\mathscr{R}_{:,S(j)_2} \in \mathbb{R}^{r_\theta \times |S(j)|}$ are required anyhow [for further explanation, see (42),(43)]. The scalings n_L^{-1}, n_R^{-1} go in hand with these to adjust the regularization terms to the magnitudes of the corresponding shares of the sampling set and disappear for the approximation of a fully available tensor (Corollary 2). The constants ζ are weighted according to the proportions of the left and right interface matrices. For example, for $\mu = 1$ (the position of the core being updated), we have $\zeta_1 = 0$ and accordingly for $\mu = 2$, $\zeta_1 \ll \zeta_2$. In practice, instead of evaluating the minimizer exactly, a few steps (that is for very coarse tolerance) of a suitably computed, *preconditioned cg* are performed as described in Sect. 5.3. This can be achieved without changing results.

Proof We omit the scalings s_1, s_2 for simplicity since they only have to be carried along the lines. We search for $N^+ := \operatorname{argmin}_{\widetilde{N}} C_{B,S,\mathscr{L},\Gamma\mathscr{N}\Theta,\mathscr{N}}(\widetilde{N})$. Substituting

$$(\Delta L, \Delta N, \Delta R) \to (\Delta\mathscr{L}\Gamma^{-1}, \Delta\mathscr{N}, \Theta^{-1}\Delta\mathscr{R})$$

we can (up to a constant factor) restate C as

$$\begin{aligned} C_{B,S,\mathscr{R},\Gamma\mathscr{N}\Theta,\mathscr{L}}(\widetilde{N}) \propto \int_{\mathbb{V}_\omega} & \|(\mathscr{L} + \Delta\mathscr{L}\Gamma^{-1})\widetilde{N}(\mathscr{R} + \Theta^{-1}\Delta\mathscr{R}) \\ & - B\|_S^2 \mathrm{d}\Delta\mathscr{L}\mathrm{d}\Delta\mathscr{N}\mathrm{d}\Delta\mathscr{R}, \\ \mathbb{V}_\omega = \{(\Delta\mathscr{L}, \Delta\mathscr{N}, \Delta\mathscr{R}) \mid & \|\Delta\mathscr{L}\|^2 + \|\Delta\mathscr{N}\|^2 + \|\Delta\mathscr{R}\|^2 \le \omega^2\}. \end{aligned}$$

Each of the independent matrices of the minimizing core is restated as

$$N^+(j) = \operatorname*{argmin}_{\widetilde{N}(j)} \int_{\mathbb{V}_\omega} \|((\mathscr{R} + \Theta^{-1}\Delta\mathscr{R})^T \otimes_K (\mathscr{L} + \Delta\mathscr{L}\Gamma^{-1}))\ \mathrm{vec}(\widetilde{N}(j)) \tag{31}$$

$$- \mathrm{vec}(B(j))\|^2_{\mathrm{vec}(S(j))} \mathrm{d}\Delta\mathscr{L}\mathrm{d}\Delta\mathscr{N}\mathrm{d}\Delta\mathscr{R}, \tag{32}$$

where $\otimes_K$ is the matrix Kronecker product. Let j be arbitrary but fixed from now on. For any x, it is $\|x\|_{\mathrm{vec}(S(j))} = \|H(j)x\|_F = x^T H(j)x$ for a diagonal, square matrix $H(j) \in \mathbb{R}^{|\mathscr{I}|/n_N \times |\mathscr{I}|/n_N}$ with $H(j)_{(s),(s)} = \delta_{s\in S(j)}$ (hence $H(j)^2 = H(j)$). Using

the normal equation, we obtain $N^+(j) = Y^{-1}b$, where

$$Y = \int_{\mathbb{V}_\omega} (\mathscr{R} + \Theta^{-1}\Delta\mathscr{R}) \otimes_K (\mathscr{L} + \Delta\mathscr{L}\Gamma^{-1})^T \\ H(j)\ (\mathscr{R} + \Theta^{-1}\Delta\mathscr{R})^T \otimes_K (\mathscr{L} + \Delta\mathscr{L}\Gamma^{-1})\mathrm{d}\Delta\mathscr{L}\mathrm{d}\Delta\mathscr{N}\mathrm{d}\Delta\mathscr{R}$$

and

$$b = \left(\int_{\mathbb{V}_\omega} (\mathscr{R} + \Theta^{-1}\Delta\mathscr{R}) \otimes_K (\mathscr{L} + \Delta\mathscr{L}\Gamma^{-1})^T\right) \\ H(j)\ \mathrm{vec}(B(j))\mathrm{d}\Delta\mathscr{L}\mathrm{d}\Delta\mathscr{N}\mathrm{d}\Delta\mathscr{R}.$$

In both Y and b, any perturbation that appears only one-sided vanishes due to symmetry of $\mathbb{V}_\omega$. Hence $b = |\mathbb{V}_\omega|\ (\mathscr{R}^T \otimes_K \mathscr{L})_{\mathrm{vec}(S(j)),:}^T\ \mathrm{vec}(B(j))_{\mathrm{vec}(S(j))}$ and for $\mathrm{d}\delta := \mathrm{d}\Delta\mathscr{L}\mathrm{d}\Delta\mathscr{N}\mathrm{d}\Delta\mathscr{R}$

$$\begin{aligned} Y = &\int_{\mathbb{V}_\omega} (\mathscr{R}^T \otimes_K \mathscr{L})^T\ H(j)\ (\mathscr{R}^T \otimes_K \mathscr{L})\mathrm{d}\delta \\ &+ \int_{\mathbb{V}_\omega} (\mathscr{R}^T \otimes_K \Delta\mathscr{L}\Gamma^{-1})^T\ H(j)\ (\mathscr{R}^T \otimes_K \Delta\mathscr{L}\Gamma^{-1})\mathrm{d}\delta \\ &+ \int_{\mathbb{V}_\omega} (\Delta\mathscr{R}^T\Theta^{-1} \otimes_K \mathscr{L})^T\ H(j)\ (\Delta\mathscr{R}^T\Theta^{-1} \otimes_K \mathscr{L})\mathrm{d}\delta \\ &+ \int_{\mathbb{V}_\omega} (\Delta\mathscr{R}^T\Theta^{-1} \otimes_K \Delta\mathscr{L}\Gamma^{-1})^T\ H(j)\ (\Delta\mathscr{R}^T\Theta^{-1} \otimes_K \Delta\mathscr{L}\Gamma^{-1})\mathrm{d}\delta \end{aligned}$$

Now, let $\ell = \#_{\mathscr{R}}$, $n = \#_{\mathscr{N}}$, $k = \#_{\mathscr{L}}$. Since $\mathbb{V}$ is a version of the $(\ell+n+k)$-sphere, we can use the following integration formula: Let $f : \mathbb{R}^{n+m} \to \mathbb{R}^k$ be a sufficiently smooth function and S_ω^{v-1} be the v-sphere of radius ω. Then

$$\int_{S_\omega^{n+m-1}} f(x_n, x_m)\mathrm{d}x = \int_0^{\pi/2} \omega \int_{S_{\omega\sin(u)}^{n-1}} \int_{S_{\omega\cos(u)}^{m-1}} f(x_n, x_m)\mathrm{d}x_m\mathrm{d}x_n\mathrm{d}u.$$

We use it twice and thereby split the integral. For a function f we then obtain

$$\begin{aligned} \int_{\mathbb{V}} f\mathrm{d}\delta &= \int_{\lambda=0}^{\omega} \int_{S_\lambda^{n+\ell+k-1}} f\mathrm{d}\delta\mathrm{d}\lambda = \int_{\lambda=0}^{\omega} \lambda \int_{g=0}^{\pi/2} \int_{S_{\lambda\sin(g)}^{n-1}} \int_{S_{\lambda\cos(g)}^{\ell+k-1}} f\mathrm{d}\delta\mathrm{d}g\mathrm{d}\lambda \\ &= \int_{\lambda=0}^{\omega} \lambda \int_{g=0}^{\pi/2} \int_{S_{\lambda\sin(g)}^{n-1}} \lambda\cos(g) \int_{u=0}^{\pi/2} \int_{S_{\lambda\cos(g)\sin(u)}^{\ell-1}} \\ &\quad \int_{S_{\lambda\cos(g)\cos(u)}^{k-1}} f\mathrm{d}\Delta\mathscr{L}\mathrm{d}\Delta\mathscr{R}\mathrm{d}u\mathrm{d}\Delta\mathscr{N}\mathrm{d}g\mathrm{d}\lambda \end{aligned}$$

If f is independent of $\Delta\mathscr{N}$, this then simplifies to

$$= \int_{\lambda=0}^{\omega} \lambda^2 \int_{g=0}^{\pi/2} |S^{n-1}_{\lambda\sin(g)}| \cos(g) \int_{u=0}^{\pi/2} \int_{S^{\ell-1}_{\lambda\cos(g)\sin(u)}} \int_{S^{k-1}_{\lambda\cos(g)\cos(u)}} f \mathrm{d}\Delta\mathscr{L}\mathrm{d}\Delta\mathscr{R}\mathrm{d}u\mathrm{d}g\mathrm{d}\lambda$$

We further use the identity (where the function $\Gamma(\cdot)$ is not to be confused with the given diagonal matrix Γ)

$$\int_0^{\pi/2} \cos(x)^p \sin(x)^q \mathrm{d}x = \frac{\Gamma((p+1)/2)\,\Gamma((q+1)/2)}{2\Gamma((p+q+2)/2)} =: \nu(p,q)$$

We apply these and Corollary 1 for different $f = (X \otimes_K Y)^T H(j)(X \otimes_K Y)$. For $\delta_1, \delta_2 \in \{0,1\}$ we set X as $\mathscr{R}^T$ ($\delta_1 = 0$) or $\Delta\mathscr{R}^T\Theta^{-1}$ ($\delta_1 = 1$) and analogously Y as $\mathscr{L}$ ($\delta_2 = 0$) or $\Delta\mathscr{L}\Gamma^{-1}$ ($\delta_2 = 1$). For the summands $Y(0,0) + Y(1,0) + Y(0,1) + Y(1,1) = Y$ this then yields

$$\begin{aligned}
Y(\delta_1,\delta_2) = &\int_{\lambda=0}^{\omega} \lambda^2 \int_{g=0}^{\pi/2} \cos(g) \frac{2\pi^{n/2}(\lambda\sin(g))^{n-1}}{\Gamma(n/2)} \\
&\int_{u=0}^{\pi/2} \frac{2\pi^{\ell/2}(\lambda\cos(g)\sin(u))^{\ell-1}}{\Gamma(\ell/2)} \left(\lambda^2\cos(g)^2\sin(u)^2\right)^{\delta_1} \\
&\frac{2\pi^{k/2}(\lambda\cos(g)\cos(u))^{k-1}}{\Gamma(k/2)} \left(\lambda^2\cos(g)^2\cos(u)^2\right)^{\delta_2} \mathrm{d}u\mathrm{d}g\mathrm{d}\lambda \cdot C_H(\delta_1,\delta_2) \\
= c \cdot &\int_{\lambda=0}^{\omega} \lambda^{n+\ell+k-1+2\delta_1+2\delta_2}\mathrm{d}\lambda \\
&\cdot \int_{g=0}^{\pi/2} \cos(g)^{\ell+k-1+2\delta_1+2\delta_2} \sin(g)^{n-1}\mathrm{d}g \\
&\cdot \int_{u=0}^{\pi/2} \cos(u)^{k-1+2\delta_2} \sin(u)^{\ell-1+2\delta_1}\mathrm{d}u \cdot C_H(\delta_1,\delta_2) \\
= c &\frac{\omega^{n+\ell+k+2\delta_1+2\delta_2}}{n+\ell+k+2\delta_1+2\delta_2} \\
&\nu(\ell+k-1+2\delta_1+2\delta_2, n-1)\ \nu(k-1+2\delta_2, \ell-1+2\delta_1)\ C_H(\delta_1,\delta_2)
\end{aligned}$$

for $c = \frac{8\pi^{(n+k+\ell)/2}}{\Gamma(n/2)\Gamma(\ell/2)\Gamma(k/2)}$. The constant matrices C_H are given by

$$\begin{aligned}
&C_H(0,0) = \widetilde{C}_H(0,0) = K(0,0)^T\ K(0,0),\ \ K(0,0) = (\mathscr{R}^T \otimes_K \mathscr{L})_{\mathrm{vec}(S(j)),:} \\
&n_L r_\gamma C_H(1,0) = \widetilde{C}_H(1,0) = K(1,0)^T\ K(1,0),\ \ K(1,0) = \mathscr{R}_{:,S(j)_2}{}^T \otimes_K \Gamma^{-1} \\
&n_R r_\theta C_H(0,1) = \widetilde{C}_H(0,1) = K(0,1)^T\ K(0,1),\ \ K(0,1) = \Theta^{-1} \otimes_K \mathscr{L}_{S(j)_1,:} \\
&|S(j)|^{-1} n_L n_R r_\gamma r_\theta C_H(1,1) = \widetilde{C}_H(1,1) = K(1,1)^T\ K(1,1),\ \ K(1,1) = \Theta^{-1} \otimes_K \Gamma^{-1}
\end{aligned}$$

Furthermore, it is $|\mathbb{V}_\omega| = c\frac{\omega^{n+\ell+k}}{n+\ell+k}\nu(\ell+k-1,n-1)\,\nu(k-1,\ell-1)$. Factoring out this base volume in $Y = |\mathbb{V}_\omega|\widetilde{Y}$ by using properties of the Γ function, one derives:

$$\widetilde{Y}(0,0) = \widetilde{C}_H(0,0), \qquad \widetilde{Y}(1,0) = n_L^{-1}\zeta_1\omega^2\widetilde{C}_H(1,0),$$
$$\widetilde{Y}(0,1) = n_R^{-1}\zeta_2\omega^2\widetilde{C}_H(0,1), \qquad \widetilde{Y}(1,1) = |S(j)|n_L^{-1}n_R^{-1}\zeta_{(1,2)}\omega^4\widetilde{C}_H(1,1),$$

where the constants s_1 and s_2 have been added again. Restating the result again as a least squares problem finishes the proof. □

Remark 6 *(Specification of constants)* Let $\#_{\mathscr{R}} := size(\mathscr{R})$, $\#_{\mathscr{N}} := size(\mathscr{N})$, $\#_{\mathscr{L}} := size(\mathscr{L})$ be the sizes of the tensor components. The constants in Theorem 3 are given by

$$\zeta_1 = s_1^2\frac{\#_{\mathscr{L}}}{r_\gamma(\#_{\mathscr{L}}+\#_{\mathscr{N}}+\#_{\mathscr{R}}+2)},$$
$$\zeta_2 = s_2^2\frac{\#_{\mathscr{R}}}{r_\theta(\#_{\mathscr{L}}+\#_{\mathscr{N}}+\#_{\mathscr{R}}+2)},$$
$$\zeta_{(1,2)} = s_1^2s_2^2\frac{\#_{\mathscr{R}}\#_{\mathscr{L}}}{r_\gamma r_\theta(\#_{\mathscr{L}}+\#_{\mathscr{N}}+\#_{\mathscr{R}}+2)(\#_{\mathscr{L}}+\#_{\mathscr{N}}+\#_{\mathscr{R}}+4)}.$$

5.3 Evaluation with coarse conjugate gradient

For each slice j, the solution to the least squares problem in Theorem 3 is described through the normal equation

$$Z(j)^T Z(j)\,\mathrm{vec}(N^+(j)) = Z(j)^T\left(B(j)|_{\mathrm{vec}(S(j))}\;0\right) \tag{33}$$

with

$$Z(j) := \begin{pmatrix}(\mathscr{R}^T\otimes\mathscr{L})|_{\mathrm{vec}(S(j)),:}\\ Y(j)\end{pmatrix},$$
$$Y(j) := \begin{pmatrix}\sqrt{n_L^{-1}\zeta_1}\;\mathscr{R}_{:,S(j)_2}{}^T\otimes\omega\Gamma^{-1}\\ \sqrt{n_R^{-1}\zeta_2}\;\omega\Theta^{-1}\otimes\mathscr{L}_{S(j)_1,:}\\ \sqrt{|S(j)|n_R^{-1}n_L^{-1}\zeta_{(1,2)}}\;\omega\Theta^{-1}\otimes\omega\Gamma^{-1}.\end{pmatrix}.$$

In practice, since within each microstep we do not benefit from an exact solution of this transitory system, it is much more economic to perform a preconditioned conjugate gradient method and terminate when a coarse, relative tolerance (e.g. $\mathrm{tol} = 10^{-2}$) is reached. This tolerance is empirically chosen such that the number of required cg steps is minimized, however under the condition that the approximation quality does not notably change in either direction—such that neither loss of accuracy nor additional regularization can be observed (cf. Sect. 8). With the following consideration, we can construct a preconditioner [cf. (14)].

Corollary 2 (Filter properties) *For full sampling, $P = \mathscr{I}$, the update is given by the so called* filter *(a diagonal matrix)*

$$\begin{aligned} \mathscr{F} &:= \big(I \otimes I + \zeta_1 \cdot I \otimes \omega^2 \Gamma^{-2} + \zeta_2 \cdot \omega^2 \Theta^{-2} \otimes I \\ &\quad + \zeta_{(1,2)} \cdot \omega^2 \Theta^{-2} \otimes \omega^2 \Gamma^{-2}\big)^{-1}, \\ \operatorname{vec}(N^+(j)) &= \mathscr{F} \operatorname{vec}\big(\mathscr{L}^T B(j) \mathscr{R}^T\big). \end{aligned} \tag{34}$$

Proof From $P = \mathscr{I}$, it follows that $\mathscr{R}_{:,S(j)_2}$ is an n_L-order copy of $\mathscr{R}$ and $\mathscr{L}_{S(j)_1,:}$ is an n_R-order copy of $\mathscr{L}$ [cf. (30)]. Hence the regularization terms are the same for all $j = 1, \ldots, n_N$. The minimizer $N^+(j)$ is given by

$$(Z(j)^T Z(j))^{-1} Z(j)^T \begin{pmatrix} \operatorname{vec}(B(j)) \\ 0 \\ \vdots \end{pmatrix}$$

$$\text{for } Z(j) = \begin{pmatrix} \mathscr{R}^T \otimes \mathscr{L} \\ \sqrt{n_L^{-1}\zeta_1}\, \mathscr{R}^T \otimes \omega \Gamma^{-1} \\ \vdots \\ \sqrt{n_R^{-1}\zeta_2}\, \omega \Theta^{-1} \otimes \mathscr{L} \\ \vdots \\ \sqrt{\zeta_{(1,2)}}\, \omega \Theta^{-1} \otimes \omega \Gamma^{-1} \end{pmatrix} \begin{matrix} \\ \\ \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} n_L\text{-times} \\ \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} n_R\text{-times} \end{matrix}.$$

The factors $\sqrt{n_L^{-1}}$ and $\sqrt{n_R^{-1}}$ vanish in $Z(j)^T Z(j)$ due to the multiple rows involving the orthogonal matrices $\mathscr{R}$ and $\mathscr{L}$. Furthermore, $(Z(j)^T Z(j))^{-1}$ is diagonal. □

Remark 7 (*Application of cg algorithm and order of computational complexity*) The matrix $\mathscr{F}$ serves as excellent preconditioner in the sense that $n_L^{-1} n_R^{-1} |S(j)| \mathscr{F}^{-1} \approx Z(j)^T Z(j)$, which holds as equality for $P = \mathscr{I}$ (as in the previous Corollary 2). This relation can as well be quantified through upper bounds on the condition number of $Z(j)\mathscr{F}^{1/2}$ as in Lemma 12 by which we expect (and observe in practice) very few iterations (and at most $r_\gamma r_\theta$) to be sufficient to reach a given coarse tolerance (for standard ALS, this is however not necessarily the case, cf. Lemma 8). Each single cg step then has complexity $\mathscr{O}(r_\gamma r_\theta |P|)$, while the full least squares problem has complexity $\mathscr{O}(r_\gamma^2 r_\theta^2 |P|)$ (per slice).

5.4 Stability and restricted isometry properties

Before we derive the central theoretical statement of this paper, Theorem 4, some preparation is necessary. The tensor restricted isometry property (e.g. [35]) does not hold for any non trivial sampling set $P \subsetneq \mathscr{I}$. We however only need to work with a modified version as follows, in which left and right interface matrices are fixed. Apart from that, the shape is exactly the same.

Definition 11 (*Internal tensor restricted isometry property (iTRIP)*) We say a rank r tensor $A = \tau_r(L, N, R)$ has the internal tensor restricted isometry property for the sampling set $S = P_{(\mu)}$, if there exist $0 \leq c < 1$ and $\rho > 0$ with

$$(1-c)\|\widetilde{A}\|_F^2 \leq \rho\|\widetilde{A}\|_S^2 \leq (1+c)\|\widetilde{A}\|_F^2$$

for all $\widetilde{A} \in \mathscr{A}(L, R) := \{\tau_r(L, \widetilde{N}, R) \mid \widetilde{N} \text{ arbitrary }\}$.

Given a tensor A, if the iTRIP does not hold, then we can not expect the next update to have good completion properties, since changes on the sampling subset are unrelated to changes on the whole space. As indicated in Example 1, if $c \to 1$, then the next iterate can be arbitrarily bad. Note that the constants only depend on the tensor A, and not on its representation, and that this property is easy to check. In particular (as proven by Lemma 12 for $\omega = 0$), the iTRIP with constant c is equivalent to

$$\kappa_2\big(\operatorname{diag}\big(\big(\mathscr{R}^T \otimes \mathscr{L}\big)\big|_{\operatorname{vec}(S(1)),:}, \ldots, \big(\mathscr{R}^T \otimes \mathscr{L}\big)\big|_{\operatorname{vec}(S(n_N)),:}\big)\big)^2 \leq \frac{1+c}{1-c},$$

where $(\mathscr{L}, N, \mathscr{R})$ corresponds to the standard representation as in (27). Hence, the property holds as long as that matrix has full rank (where κ_2 is the condition number regarding the spectral norm $\|\cdot\|_2$) or equivalently $N \mapsto (LNR)_S$ is injective. Note that, by a slice wise consideration, the condition number of a single $(\mathscr{R}^T \otimes \mathscr{L})|_{\operatorname{vec}(S(j)),:}$ can be improved since different magnitudes of sampling for each slice can be compensated (after all, the slices are solved independently).

Lemma 8 (Likelihood of the iTRIP) *Let $\mathscr{T}_r$ be the subset of 3 dimensional tensors with rank $r = (r_\gamma, r_\theta)$. Let P be a (random) sampling that fulfills $|S(j)| \geq r_\gamma r_\theta$ for all $j = 1, \ldots, n_N$. Then, for (only) almost every representation $(L, N, R) \in \mathbb{R}^{n_L \times r_\gamma} \times \mathbb{R}^{r_\gamma \times n_N \times r_\theta} \times \mathbb{R}^{r_\theta \times n_R}$ (with respect to the Lebesgue measure), the tensor $A = \tau_r(L, N, R) \in \mathscr{T}_r$ has the iTRIP. If for one j, $|S(j)| < r_\gamma r_\theta$, then no $A \in \mathscr{T}_r$ has the iTRIP.*

Proof A tensor $A = \tau_r(L, N, R)$ has the iTRIP (for some valid constants) if and only if the linear map $N \mapsto (LNR)_S$ is injective, or equivalently, $(R^T \otimes L)_{\operatorname{vec}(S(j)),:}$ has full rank for each j. Due to the provided slice density of P, each matrix $(\mathscr{R}^T \otimes \mathscr{L})_{\operatorname{vec}(S(j)),:}$ is of size $|S(j)| \times r_\gamma r_\theta$. Hence generically, it is of full rank. If $|S(j)| < r_\gamma r_\theta$, then the matrix cannot have full rank. □

Tensors themselves that do not have the iTRIP, assuming sufficient sampling, pose just a marginal phenomenon for high dimension d. A quite simple construction however shows that the iTRIP does not behave well under perturbation:

Lemma 9 (iTRIP under perturbation) *Let $B \in \mathbb{R}^{n_L \times n_N \times n_R}$ with singular values $(\gamma^{(B)}, \theta^{(B)})$. Assume further that for one j it holds $|S(j)_1| < n_L$. Then for every $\sigma^* > 0$, there exists a tensor A with rank (r_γ, r_θ) and $\|A - B\|_F^2 \leq \sum_{i=r_\gamma}^{\infty} (\gamma_i^{(B)})^2 + 2\sum_{i=r_\theta+1}^{\infty} (\theta_i^{(B)})^2 + (\sigma^*)^2$ such that A does not have the iTRIP (and $\gamma_{r_\gamma}^{(A)} = \sigma^*$). If B already has rank (r_γ, r_θ), then $\|A - B\|_F^2 \leq (\gamma_{r_\gamma}^{(B)})^2 + (\sigma^*)^2$ suffices.*

Proof Truncation of B yields a tensor $\widetilde{A}$ with rank $(*, r_\theta)$ and $\|\widetilde{A} - B\|_F^2 \leq \delta := \sum_{i=r_\theta+1}^{\infty}(\theta_i^{(B)})^2$. The tensor $\widetilde{A}$ hence has perturbed singular values such that $\|\gamma^{(\widetilde{A})} - \gamma^{(B)}\|_2 \leq \delta$ (Mirsky's Theorem [30]). Let $(\widetilde{\mathscr{L}}, \Gamma^{(\widetilde{A})}, \widetilde{\mathscr{N}}, \Theta^{(\widetilde{A})}, \widetilde{\mathscr{R}})$ be the standard representation of $\widetilde{A}$. Without loss of generality, we may assume that $j = 1$ and that only points in the first $k := |S(1)|$ rows of $B(1)$ are contained in the sampling $S(1)$. Let now

$$\widetilde{\mathscr{L}}_{:,\{1,\dots,r_\gamma\}} =: \begin{pmatrix} X & \widetilde{x} \\ Y & \widetilde{y} \end{pmatrix}, \quad \mathscr{L} := \begin{pmatrix} X & x \\ Y & y \end{pmatrix}, \quad X \in \mathbb{R}^{k \times r_\gamma - 1},$$

If X is already singular, then we may choose $A = \widetilde{A}$. Otherwise, then we may choose $x = \alpha X v$, $y = \alpha \widehat{y}$ for $\alpha = \|(Xv; \widehat{y})\|_2^{-1}$ and $v = -(X^T X)^{-1} Y^T \widehat{y}$, for an arbitrary vector $\widehat{y} \neq 0$. In all three cases, $\mathscr{L}$ is orthogonal and for $A := \tau_r(\mathscr{L}, \mathrm{diag}(\gamma_1^{(\widetilde{A})}, \dots, \gamma_{r_\gamma - 1}^{(\widetilde{A})}, \sigma^*)\widetilde{\mathscr{N}}\Theta^{(\widetilde{A})}, \widetilde{\mathscr{R}})$ it holds $\|A - B\|_F \leq \|A - \widetilde{A}\|_F + \|\widetilde{A} - B\|_F \leq \sum_{i=r_\gamma}^{\infty}(\sigma_\gamma^{(\widetilde{A})})_i^2 + (\sigma^*)^2 + \delta \leq \sum_{i=r_\gamma}^{\infty}(\sigma_\gamma^{(B)})_i^2 + \delta + (\sigma^*)^2 + \delta$. Yet $(\mathscr{R}^T \otimes \mathscr{L})|_{\mathrm{vec}(S(1)),:}$ is a singular matrix, since $\mathscr{L}_{\{1,\dots,k\},:}$ is already singular. □

The statement analogously holds true for r_θ and can easily be transferred to matrix completion as well. Tensors that do not have the iTRIP are hence densely scattered depending on $\gamma_{r_\gamma}^{(B)}$, as are, more importantly, surroundings in which the constant c is close to 1 and overfitting becomes more likely (cf. Example 1). In case of the regularized update, the additional term $Y(j)$ (30) does not allow the condition number to change that easily (cf. Lemma 11).

Lemma 10 (Partial matrix inverse by divergent parts) *We partition* $\{1, \dots, n\} = \omega_j \cup \omega_j^c$ $(\omega_j^c = \{1, \dots, n\} \setminus \omega_j)$, $j = 1, 2$ *and define* $\Omega := \omega_1 \times \omega_2$, $\widetilde{\Omega} := \omega_1^c \times \omega_2^c$. *Let* $\{A^{(k)}\}_k$, $\{J^{(k)}\}_k \subset \mathbb{R}^{n \times n}$ *be series of symmetric matrices,* $supp(J^{(k)}) \subset \Omega$.
If $lim_{k\to\infty} A^{(k)}|_{\widetilde{\Omega}} = A|_{\widetilde{\Omega}}$, $A|_{\widetilde{\Omega}}$ *s.p.d, and* $\sigma_{\min}(J^{(k)}|_\Omega) \to \infty$, *then* $V := lim_{k\to\infty}(A^{(k)} + J^{(k)})^{-1}$ *exists and we have* $V|_{\widetilde{\Omega}} = (A|_{\widetilde{\Omega}})^{-1}$ *and* $V|_{\widetilde{\Omega}^c} = 0$ $(\widetilde{\Omega}^c = \{1, \dots, n\}^2 \setminus \widetilde{\Omega})$.

Proof First, w.l.o.g., let $\Omega = \{m+1, \dots, n\}^2$. Otherwise we can apply permutations. Further, let $V^{(k)} := A^{(k)} + J^{(k)}$. We partition our (symmetric) matrices M for $M_{1,1} \in \mathbb{R}^{m \times m}$ block-wise as

$$M = \begin{pmatrix} M_{1,1} & M_{1,2} \\ M_{1,2}^T & M_{2,2} \end{pmatrix}.$$

Note that $J_{1,1}^{(k)}, J_{1,2}^{(k)} \equiv 0$. Since $A_{1,1}^{(k)} = V_{1,1}^{(k)}$ and $A_{1,1} = A|_{\widetilde{\Omega}}$ is s.p.d, $A_{1,1}^{(k)}$ is invertible for all $k > K$ for some K and hence $\lim_{k\to\infty}(V_{1,1}^{(k)})^{-1} = A_{1,1}^{-1}$. Further, $\sigma_{\min}(B_{2,2}^{(k)}) > \sigma_{\min}(J_{2,2}^{(k)}) - \sigma_{\max}(A_{2,2}^{(k)}) \to \infty$ and hence $\|(V_{2,2}^{(k)})^{-1}\| \to 0$. Therefore, for $k > \widetilde{K}$ and $H^{(k)} := V_{1,1}^{(k)} - V_{1,2}^{(k)}(V_{2,2}^{(k)})^{-1}(V_{1,2}^{(k)})^T$, it is $\sigma_{\min}(H^{(k)}) > \sigma_{\min}(A_{1,1})/2$. By block-wise inversion of $V^{(k)}$, it then follows $((V^{(k)})^{-1})_{1,1} = (H^{(k)})^{-1} \to (A_{1,1}^{(k)})^{-1}$. Similarly, $((V^{(k)})^{-1})|_\Omega \to 0$. □

One last step remains, since we cannot allow ζ to depend on the rank r. For now, we redefine the method $\mathscr{M}^*$ to directly yield the result in Theorem 3 for arbitrary constants ζ, i.e.

$$\mathscr{M}^*_\zeta(L,\ N,\ R) := (\mathscr{L},\ N^+,\ \mathscr{R}). \tag{35}$$

We explain in Sect. 7 and Lemma 12 how the scalings s_1, s_2 as well as ω are used to obtain one specific $\mathscr{M}^*_\zeta$ from $\mathscr{M}^*$, for which ζ is indeed independent of r.

Theorem 4 (Stability of the method $\mathscr{M}^*_\zeta$) *Let B be the target tensor, $S \subsetneq \mathscr{I}$ the sampling set, arbitrary but fixed, and $\mathscr{M}^*_\zeta$ as in (35).*

- *The regularized method $\mathscr{M}^*_\zeta$ ($\omega > 0$) as defined by (35) (for $\zeta_1, \zeta_2 \geq 0$ and $\zeta_{(1,2)} > 0$ that do not depend on r) is stable at all points A^* (and hence also fixed-rank stable).*
- *The unregularized method ($\omega = 0$) (23) provides stability only for fixed rank (cf. Example 2), and only at those points A^* that have the iTRIP (cf. Def. 11).*

Proof Let A^* be a fixed tensor with TT-ranks r^*.

1. *Fixed-rank stability:* We first show that $\mathscr{M}^*$ is stable for fixed rank. Let A_i be a sequence with $\operatorname{rank}(A_i) = r^*$ and $A_i \to A^*$. Let $\mathscr{G}^* = (\mathscr{L}^*, \Gamma^*, \mathscr{N}^*, \Theta^*, \mathscr{R}^*)$ be the standard representation of A^* as well as $\mathscr{G}_i$ correspond to A_i. We partition the indices for γ^* and θ^* by k and ℓ according to equality of entries, such that $\gamma^*_1 = \ldots = \gamma^*_{k_1} > \gamma^*_{k_1+1} = \ldots = \gamma^*_{k_2} > \ldots > \gamma^*_{k_{K-1}+1} = \ldots = \gamma^*_{k_K} > 0$ and likewise for $\ell_1, \ldots, \ell_L$. Since $A_i \to A^*$, their singular values also converge (e.g. [44]). We can hence conclude from [11,42] that there exist sequences of block diagonal, orthogonal matrices W_i and M_i with block sizes $k_1, k_2 - k_1, \ldots, k_K - k_{K-1}$ and $\ell_1, \ell_2 - \ell_1, \ldots, \ell_L - \ell_{L-1}$, respectively, such that

$$\|\mathscr{L}_i W_i - \mathscr{L}^*\|_F \to 0 \quad \text{and} \quad \|M_i \mathscr{R}_i - \mathscr{R}^*\|_F \to 0, \tag{36}$$

since the standard representation includes left and right singular vectors. We have to show that the tensors $H_i = \tau_r(\mathscr{L}_i, N_i, \mathscr{R}_i) = \tau_r(\mathscr{M}^*(\mathscr{L}_i, \Gamma_i \mathscr{N}_i \Theta_i, \mathscr{R}_i))$ converge to the analogously defined H^*. For fixed j, we define for each single $\mathscr{G}_i$ the matrix $Y_i = Y(j)$ (cf. Theorem 3, Remark 33) and $z_i := (\mathscr{R}_i^T \otimes \mathscr{L}_i)$ such that

$$N_i(j) = \operatorname*{argmin}_{\widetilde{N}(j)} \left\| \begin{pmatrix} (z_i)_{\operatorname{vec}(S(j)),:} \\ Y_i \end{pmatrix} \operatorname{vec}(\widetilde{N}(j)) - \begin{pmatrix} \operatorname{vec}(B(j))|_{\operatorname{vec}(S(j))} \\ 0 \end{pmatrix} \right\|,$$
$$\operatorname{vec}(H_i(j)) = z_i \operatorname{vec}(N_i(j)). \tag{37}$$

We define the shifted matrices

$$z_i^{M,W} := (M_i\ \mathscr{R}_i)^T \otimes (\mathscr{L}_i\ W_i)$$
$$Y_i^{M,W} := \begin{pmatrix} \sqrt{n_L^{-1}\zeta_1}\ (M_i\ \mathscr{R}_{:,S(j)_2})^T \otimes (\Gamma_i^{-1}\ W_i) \\ \sqrt{n_R^{-1}\zeta_2}\ (\Theta_i^{-1}\ M_i^T) \otimes (\mathscr{L}_{S(j)_1,:}\ W_i) \\ \sqrt{|S(j)| n_R^{-1} n_L^{-1} \zeta_{(1,2)}}\ (\Theta_i^{-1}\ M_i^T) \otimes (\Gamma_i^{-1}\ W_i) \end{pmatrix}$$

Due to (36), it holds $(z_i^{M,W})_{\mathrm{vec}(S(j)),:} \to z^*_{\mathrm{vec}(S(j)),:}$. Inserting $I = (M_i^T \otimes W_i)(M_i^T \otimes W_i)^T$ into (37), we obtain

$$\mathrm{vec}(H_i(j)) = z_i^{M,W} \left(\left(z_i^{M,W}\right)^T_{\mathrm{vec}(S(j)),:} \left(z_i^{M,W}\right)_{\mathrm{vec}(S(j)),:} + Y_i^{M,W^T} Y_i^{M,W} \right)^{-1} \cdot \left(z_i^{M,W}\right)^T_{\mathrm{vec}(S(j)),:} \mathrm{vec}(B(j))|_{\mathrm{vec}(S(j))}.$$

Since $W_i^T \Gamma^* W_i = \Gamma^*$ for all i, it follows $W_i^T \Gamma_i W_i \to \Gamma^*$. Likewise $M_i^T \Theta_i M_i \to \Theta^*$ and thereby also $Y_i^{M,W^T} Y_i^{M,W} \to Y^{*T} Y^*$. We treat the cases $\omega = 0$ and $\omega > 0$ separately:

(i) $\omega = 0$: In this case, $Y_i^{M,W} = 0 = Y^*$. If the iTRIP holds for A^*, then $\sigma_{\min}(z^*_{\mathrm{vec}(S(j)),:}) > 0$ and therefore

$$\left(\left(z_i^{M,W}\right)^T_{\mathrm{vec}(S(j)),:} \left(z_i^{M,W}\right)_{\mathrm{vec}(S(j)),:} \right)^{-1} \to \left((z^*)^T_{\mathrm{vec}(S(j)),:} (z^*)_{\mathrm{vec}(S(j)),:} \right)^{-1}.$$

This directly yields convergence of $(H_i(j)) \to (H^*(j))$ since all involved factors converge.

(ii) $\omega > 0$: Here, we use that $\sigma_{\min}(Y^*) > 0$ and $\sigma_{\min}(z^*_{\mathrm{vec}(S(j)),:}) \geq 0$. We then obtain convergence since

$$\left(\left(z_i^{M,W}\right)^T_{\mathrm{vec}(S(j)),:} \left(z_i^{M,W}\right)_{\mathrm{vec}(S(j)),:} + Y_i^{M,W^T} Y_i^{M,W} \right)^{-1} \to \left((z^*)^T_{\mathrm{vec}(S(j)),:} (z^*)_{\mathrm{vec}(S(j)),:} + Y^{*T} Y^* \right)^{-1}.$$

This proves fixed-rank stability.

2. *Stability:* Let now A_i have arbitrary ranks. Without loss of generality by consideration of a finite amount of infinite subsequences, we can assume that $\mathrm{rank}(A_i) \equiv r$ for all i. Then, since $TT(r^*)$ is a manifold, it follows $\gamma \geq \gamma^*$ and $\theta \geq \theta^*$. We can therefore have singular values $(\gamma_i)_{k_K+1}, \ldots, (\gamma_i)_{k_{K+1}} \to 0$ as well as $(\theta_i)_{\ell_L+1}, \ldots, (\theta_i)_{\ell_{L+1}} \to 0$. We expand the matrices W_i and M_i by identities of appropriate sizes to account for the vanishing singular values: $W_i \leftarrow diag(W_i, I_{k_{K+1}-k_K})$, $M_i \leftarrow diag(M_i, I_{\ell_{L+1}-\ell_L})$. In regard of Proposition 10, let Ω be the smallest cross product set, such that $(Y_i^{M,W^T} Y_i^{M,W})|_{\widetilde{\Omega}}$ converges (which is the set that corresponds to vanishing singular values). Then, due to the definition of $Y_i^{M,W}$, $\sigma_{\min}((Y_i^{M,W^T} Y_i^{M,W})|_{\Omega}) \to \infty$. We can conclude that

$$\left(\left(\left(z_i^{M,W}\right)^T_{\mathrm{vec}(S(j)),:} \left(z_i^{M,W}\right)_{\mathrm{vec}(S(j)),:} + Y_i^{M,W^T} Y_i^{M,W} \right)^{-1} \right)\Big|_{\widetilde{\Omega}} \to \left((z^*)^T_{\mathrm{vec}(S(j)),:} (z^*)_{\mathrm{vec}(S(j)),:} + Y^{*T} Y^* \right)^{-1}.$$

and

$$\left(\left(\left(z_i^{M,W}\right)^T_{\mathrm{vec}(S(j)),:}\left(z_i^{M,W}\right)_{\mathrm{vec}(S(j)),:} + Y_i^{M,W^T}\, Y_i^{M,W}\right)^{-1}\right)\Big|_{\widetilde{\Omega}^c} \to 0.$$

Because of this restriction, we in turn again get convergence to the limit $(H_i(j)) \to (H^*(j))$, since all parts that correspond to vanishing singular values, also vanish within the update. This finishes the proof. □

Definition 12 (*Stabilized internal tensor restricted isometry property (siTRIP)*) We say a rank r tensor $A = \tau_r(L, N, R)$ has the stable internal tensor restricted isometry property for the sampling set S, if there exist $0 \le c < 1$ and $\rho > 0$ such that for all $\widetilde{N}$ holds

$$\begin{aligned}(1-c)\int_{\mathscr{V}_\omega(L,N,R)} \|\widetilde{A}_\Delta\|_F^2 \mathrm{d}\Delta &\le \rho \int_{\mathscr{V}_\omega(L,N,R)} \|\widetilde{A}_\Delta\|_S^2 \mathrm{d}\Delta \\ &\le (1+c)\int_{\mathscr{V}_\omega(L,N,R)} \|\widetilde{A}_\Delta\|_F^2 \mathrm{d}\Delta \end{aligned} \tag{38}$$

where $\widetilde{A}_\Delta := \tau_r(L+\Delta L, \widetilde{N}+\Delta N, R+\Delta R)|_F^2$ and $\mathrm{d}\Delta = \mathrm{d}\Delta L \mathrm{d}\Delta N \mathrm{d}\Delta R$.

The constants are independent of the specific representation (cf. Lemma 7).

Lemma 11 *Let* $Z(j) = \begin{pmatrix} (\mathscr{R}^T \otimes \mathscr{L})|_{\mathrm{vec}(S(j)),:} \\ Y(j) \end{pmatrix}$ *as in* (33). *The siTRIP with constant* c *for* $A \in \mathbb{R}^{\mathscr{I}}$ *is equivalent to*

$$\exists\, c > 0: \quad \kappa_2(\mathrm{diag}(Z(1)\mathscr{F}^{1/2}, \ldots, Z(n_N)\mathscr{F}^{1/2})^2 \le \frac{1+c}{1-c},$$

where $A = \tau_r(\mathscr{L}, N, \mathscr{R})$ *is its standard representation and* $\mathscr{F}$ *is as in Lemma 2.*

Proof Let $Z = \mathrm{diag}(Z(1), \ldots, Z(n_N))$ and $\mathrm{vec}(\widetilde{N})^T = (\mathrm{vec}(N(1))^T, \ldots, \mathrm{vec}(N(n_N))^T)$. We abbreviate the siTRIP (38) as $(1-c)\beta \le \rho\xi \le (1+c)\beta$ ($\beta = \beta(\widetilde{N})$, $\xi = \xi(\widetilde{N})$). Let $a := \max_{x\neq 0} \frac{\|Zx\|^2}{\|(I\otimes\mathscr{F}^{-1/2})x\|^2}$ and $b := \min_{x\neq 0} \frac{\|Zx\|^2}{\|(I\otimes\mathscr{F}^{-1/2})x\|^2}$. Since the perturbation ΔN is independent of $\widetilde{N}$, this term can be neglected in consideration of that $\|\widetilde{N}\|$ is not bounded. With $P = \mathscr{I}$ it holds $\beta = \sum_{j=1}^{n_N} \mathrm{vec}(\widetilde{N}(j))^T \mathscr{F}^{-1} \mathrm{vec}(\widetilde{N}(j)) = \|(I\otimes\mathscr{F}^{-1/2})\mathrm{vec}(\widetilde{N})\|_F^2$ (cf. proof of Theorem 3). For the actual sampling P, we have $\xi = \|Z\mathrm{vec}(\widetilde{N})\|_2^2$. Thereby $a = \max_{\widetilde{N}|\beta=1} \xi$ and $b = \min_{\widetilde{N}|\beta=1} \xi$. Now, given the siTRIP, it follows

$$\mathrm{cond}\big(Z\big(I\otimes\mathscr{F}^{1/2}\big)\big)^2 = \frac{a}{b} \le \frac{\rho^{-1}(1+c)}{\rho^{-1}(1-c)} = \frac{1+c}{1-c}.$$

For the opposite implication, define $\rho = \frac{1-c}{b}$. Then

$$\rho\xi \le a\frac{1-c}{b}\beta \le (1+c)\beta \quad \text{and} \quad \rho\xi \ge b\frac{1-c}{b}\beta = (1-c)\beta.$$

□

The siTRIP holds for *every* tensor (possibly with c close to 1). For $\omega \to 0$, the constant c converges to the one of the iTRIP and for $\omega \to \infty$, $c \to 0$. As well as for Definition 11, a slice wise consideration yields a better condition number for just $Z(j)\mathscr{F}^{1/2}$. Note that this value not only bounds the number of steps required for the cg method, but appears to be important for the reconstruction quality obtained through one microstep. However, a further investigation into the siTRIP, how exactly it behaves under perturbation and if it requires modifications, is a matter of future research.

As for the matrix case, we have to limit the singular values from below by a decreasing value proportional to the current residual, cf. Sect. 6. This leads to a slight complication, which is resolved through the following, simple corollary to Theorem 3.

Corollary 3 *Let A and $\widetilde{A}$ be tensors with standard representations $(\mathscr{L}, \Gamma, \mathscr{N}, \Theta, \mathscr{R})$ and $(\mathscr{L}, \Gamma, \widetilde{\mathscr{N}}, \Theta, \mathscr{R})$, respectively. Then both yield the same update N^+.*

Hence, if we want to modify the singular values γ and θ, we may do so without knowledge about an appropriate core $\widetilde{\mathscr{N}}$ (for example in the sense of some unknown best approximation). In the least squares problem in Theorem 3, we therefor simply set $\gamma_i := \max(\gamma_i, \sigma_{\min}), i = 1, \dots, r_{\mu-1}$ and $\theta_i := \max(\theta_i, \sigma_{\min}), i = 1, \dots, r_\mu$ (and can thereby also ignore that the combination of the new γ and θ might not be feasible, cf. [26]

6 Behavior of the SALSA filter

A deeper understanding of the regularization utilized by SALSA and the reason for the lower bound $\sigma_{\min}$ is provided by the *filter* as indicated by (14) for the matrix case and as defined by Corollary 2 for tensors. Throughout this section, we assume that the sampling is such that for the minimizer in Theorem 3 it (approximately) holds

$$\operatorname{vec}(N^+(j)) = \mathscr{F}\ \operatorname{vec}\big(\big(\mathscr{L}^T\ B(j)\ \mathscr{R}^T\big)\big), \tag{39}$$

which is true at last for $P = \mathscr{I}$ (cf. Corollary 2). Since $\zeta_1\zeta_2 = \zeta_{(1,2)}$ [see the later equation (41)], we can rewrite

$$\begin{aligned} N^+ &= D_{\omega^2\zeta_1}(\Gamma)\ \big(\mathscr{L}^T\ B\ \mathscr{R}^T\big)\ D_{\omega^2\zeta_2}(\Theta) \\ D_c(\Sigma) &:= (I + c\Sigma^{-2})^{-1}. \end{aligned}$$

We are interested in the fixpoints of this update, i.e. we postulate $N^+ = \Gamma\ \mathscr{N}\ \Theta$. Then, since $\mathfrak{R}(\mathscr{N}\ \Theta)$ is row-orthogonal (cf. Lemma 6), it holds

$$\begin{aligned} D_{\omega^2\zeta_1}(\Gamma)\ Z &= \Gamma, \\ Z &= \mathfrak{R}\big(\big(\mathscr{L}^T\ B\ \mathscr{R}^T\big)\ D_{\omega^2\zeta_2}(\Theta)\big)\ \mathfrak{R}(\mathscr{N}\ \Theta)^T, \end{aligned} \tag{40}$$

where $Z =: \operatorname{diag}(\sigma^{(Z)})$ is necessarily a diagonal matrix (certainly, an analogous argument holds for Θ as well). Because (40) can only hold if $d_{\sigma^{(Z)},\zeta_1}(\gamma_i) = \gamma_i$

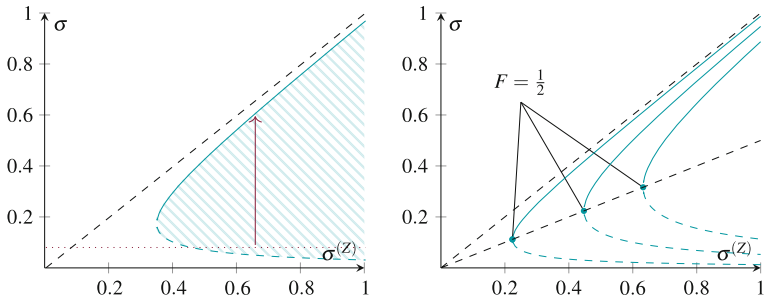

Fig. 4 Left: Plotted are the fixpoints (continuous for attractive, dashed for repelling ones, in teal) of $d_{\sigma^{(Z)},c}$ for one fixed c with respect to $\sigma^{(Z)}$. Within the hatched area, singular values rise until they reach the upper boundary. A lower limit to the singular values is indicated as dotted, magenta line. Right: Different values of c are considered. The turning point $\sigma = c = \frac{1}{2}\sigma^{(Z)}$ corresponds to a filter value of $1/2$

for all i, the focus of our analysis lies on the fixpoints of the function $d_{\sigma^{(Z)},c} : \sigma \mapsto (1 + c\sigma^{-2})^{-1}\sigma^{(Z)}$. For each pair $(\sigma^{(Z)}, c)$, the only attractive fixpoint (if existent) is given by $f_{\text{stab}} = \frac{1}{2}\sigma^{(Z)} + \frac{1}{2}\sqrt{(\sigma^{(Z)})^2 - 4c}$ and the repelling one by $f_{\text{rep}} = \frac{1}{2}\sigma^{(Z)} - \frac{1}{2}\sqrt{(\sigma^{(Z)})^2 - 4c}$. At the point where $f_{\text{stab}} = f_{\text{rep}}$, it holds $\sigma = \sqrt{c} = \frac{1}{2}\sigma^{(Z)}$. The minimal value which the term $(1 + c\sigma^{-2})^{-1}$ can hence take in any attractive fixpoint, is $F = 1/2$. This behavior is shown in Fig. 4. A stabilized singular value corresponds to some attractive fixpoint of $d_{\sigma^{(Z)},c}$ (cf. Definition 4), and it necessarily holds $(D_{\omega^2\zeta_1}(\Gamma))_{i,i} > 0.5 \Leftrightarrow \gamma_i > \omega\sqrt{\zeta_1}$. This explains why the lower limit $\sigma_{\min}$ is necessary. As displayed in Fig. 4 (left), for any fixed $\sigma^{(Z)}$, a singular value σ must be above a certain threshold (that corresponds to the repelling fixpoint) to be increased by the microstep. It therefore must not converge to zero.

7 Results transferred back to a d-dimensional tensor

In this Section, we transfer the previous results for $S = P_{(\mu)}$ and $B = M_{(\mu)}$ to the d-dimensional setting. In Remark 6, we have $\#_{\mathscr{R}} = size(\mathscr{R}) = r_\theta \prod_{i=s+1}^{d} n_i$, $\#_{\mathscr{N}} = size(\mathscr{N}) = r_\gamma n_s r_\theta$, $\#_{\mathscr{L}} = size(\mathscr{L}) = r_\gamma \prod_{i=1}^{s-1} n_i$. By combining modes (cf. Definition 4), the *sizes* of the left as well as right side have been distorted, considering that the degrees of freedom of $\mathscr{L} = \mathscr{G}^{<s}$ and $\mathscr{R} = \mathscr{G}^{>s}$ are given by a sum, not a product, of the degrees of freedom of the single modes (ignoring minor gauge conditions). We choose one of the few remaining options through which the method becomes stable. We artificially set

$$\#_{\mathscr{R}} \leftarrow r_\mu \sum_{i=\mu+1}^{d} n_i, \quad \#_{\mathscr{L}} \leftarrow r_{\mu-1} \sum_{i=1}^{\mu-1} n_i$$

using appropriate scalings $s_1 = s_1^{(\mu)}$, $s_2 = s_2^{(\mu)}$ (differently for each mode μ). Otherwise, we will not obtain a stable microstep. Furthermore, the near common parts of

the denominators, $\#_{\mathscr{R}} + \#_{\mathscr{N}} + \#_{\mathscr{R}} + 2(+2)$, can be incorporated into ω^2, so we omit them in the following sense:

Lemma 12 (Rescaled target function) *The previously discussed rescaling is achieved by choosing*

$$(s_1^{(\mu)})^2 = E\frac{\sum_{s=1}^{\mu-1} n_s}{\left(\prod_{s=1}^{\mu-1} n_s\right)\sum_{s=1}^{d} n_s}, \quad (s_2^{(\mu)})^2 = E\frac{\sum_{s=\mu+1}^{d} n_s}{\left(\prod_{s=\mu+1}^{d} n_s\right)\sum_{s=1}^{d} n_s},$$
$$E = r_\mu \prod_{s=\mu+1}^{d} n_s + r_{\mu-1} n_\mu r_\mu + r_{\mu-1}\prod_{s=1}^{\mu-1} n_s$$

Thereby,

$$\zeta_1^{(\mu)} = \frac{\sum_{s=1}^{\mu-1} n_s}{\sum_{s=1}^{d} n_s}, \quad \zeta_2^{(\mu)} = \frac{\sum_{s=\mu+1}^{d} n_s}{\sum_{s=1}^{d} n_s}, \quad \zeta_{(1,2)}^{(\mu)} = \zeta_1^{(\mu)}\zeta_2^{(\mu)}(1 + \mathscr{O}(E^{-1})). \tag{41}$$

Proof First, $\zeta_1^{(\mu)} = s_1^2 \frac{\prod_{s=1}^{\mu-1} n_s}{E} = \frac{\sum_{s=1}^{\mu-1} n_s}{\sum_{s=1}^{d} n_s}$, with an analog result for $\zeta_2^{(\mu)}$. For the mixed term, we have $\zeta_{(1,2)}^{(\mu)} = \zeta_1^{(\mu)}\zeta_2^{(\mu)}\frac{E}{E+2} = \zeta_1^{(\mu)}\zeta_2^{(\mu)}(1 - \frac{2}{E+2})$. □

The value E^{-1} is in general far below machine accuracy, such that we (from now on) ignore the factor $(1 + \mathscr{O}(E^{-1}))$. There might be a more suitable realization of this result and it should be remarked that the exact scalings are not important for the validity of Theorem 4. In this context, for fixed μ, the matrices $\mathscr{L}_{S(j)_1,\cdot\cdot} \in \mathbb{R}^{|P_{(\mu)}(j)|\times r_\gamma}$ and $\mathscr{R}_{\cdot,S(j)_2} \in \mathbb{R}^{r_\theta\times|P_{(\mu)}(j)|}$ [cf. (22)], are given by

$$\left(\mathscr{L}_{S(j)_1,\cdot\cdot}\right)_{\ell,\cdot\cdot} = G_1(p_1^{(i_\ell)}) \cdot \ldots \cdot G_{\mu-1}\left(p_{\mu-1}^{(i_\ell)}\right) = G_{1,\ldots,\mu-1}\left(\left(p_1^{(i_\ell)},\ldots,p_{\mu-1}^{(i_\ell)}\right)\right), \tag{42}$$

$$\left(\mathscr{R}_{\cdot,S(j)_2}\right)_{\cdot,\ell} = G_{\mu+1}\left(p_{\mu+1}^{(i_\ell)}\right) \cdot \ldots \cdot G_d(p_d^{(i_\ell)}) = G_{\mu+1,\ldots,d}\left(\left(p_{\mu+1}^{(i_\ell)},\ldots,p_d^{(i_\ell)}\right)\right), \tag{43}$$

for $p^{(i_\ell)} \in P_{(\mu)}(j)$, $\ell = 1,\ldots,|P_{(\mu)}(j)|$ and a representation G for which $\mathscr{L} = G^{<s}$ and $\mathscr{R} = G^{>s}$ [cf. (21)].

Remark 8 (*Case* $\mu = 1, d$) For $\mu = 1, d$ in Theorem 3, the same formula can be used by formally setting $G^{<1} = \mathscr{L} = 1$, $G^{>d} = \mathscr{R} = 1$ and $\zeta_1^{(1)} = 0$, $\zeta_2^{(d)} = 0$, $\zeta_{(1,2)}^{(1)}, \zeta_{(1,2)}^{(d)} = 0$, respectively. These comply with the result in the matrix case (cf. Remark 2).

Since all microsteps $\mathscr{M}^*$ are stable, we call this regularized ALS method stable—hence the name SALSA (Stable ALS Approximation). We summarize in Algorithm 3 one full left sweep $\mu = 1 \to d$ of SALSA for fixed rank r. Note that in practice, the complexity is reduced to the minimal necessary order in the optimal case (cf. Remark 7). The simpler matrix case ($d = 2$) is carried out in Algorithm 1.

Algorithm 3 SALSA Sweep

we here identify $\widetilde{\Sigma} = \mathrm{diag}(\widetilde{\sigma})$

Require: limit $\sigma_{\min}$, parameter ω, initial guess $A = \tau_r(G)$ for which $\mathfrak{R}(G_2), \ldots, \mathfrak{R}(G_d)$ are row-orthogonal and data points $M|_P$

for $\mu = 1, \ldots, d$ **do**
  if $\mu \neq 1$ **then**
    compute the SVD $U\widetilde{\Sigma}V^T := \mathfrak{L}(G_{\mu-1})$ and set $\sigma_i^{(\mu-1)} := \max(\widetilde{\sigma}_i, \sigma_{\min}), i = 1, \ldots, r_{\mu-1}$
    set $G_{\mu-1}$ via $\mathfrak{L}(G_{\mu-1}) = U$ and $G_\mu := \widetilde{\Sigma}V^T G_\mu$
  end if
  if $\mu \neq d$ **then**
    compute the SVD $U\widetilde{\Sigma}V^T := \mathfrak{L}(G_\mu)$ and set $\sigma_i^{(\mu)} := \max(\widetilde{\sigma}_i, \sigma_{\min}), i = 1, \ldots, r_\mu$
    update $G_{\mu+1} := V^T G_{\mu+1}$ and G_μ via $\mathfrak{L}(G_\mu) = U\widetilde{\Sigma}$
  end if
  for $j = 1, \ldots, n_\mu$ **do**
    update $G_\mu(j) := N(j)$ by solving the least squares problem in Theorem 3 for $\mathscr{L} = G^{<s}$, $\mathscr{R} = G^{>s}$, $\gamma = \sigma^{(\mu-1)}$, $\theta = \sigma^{(\mu)}$ (cf. Remark 8) using coarse cg (cf. Remark 7)
  end for
end for

8 Semi implicit and non uniform rank adaption

The rank adaption for tensor completion is carried out analogously to the matrix case, Sect. 3.2. The exact choices of the following parameters are not important, such that we only indicate them roughly. The specific values which we used in all numerical tests are provided in Sect. 9.4. The number of minor singular values (cf. Definition 4) for each matricization is kept constant, such that

$$\left|\left\{i \mid 0 < \sigma_i^{(\mu)} < f_{\text{minor}} \cdot \omega\right\}\right| \stackrel{!}{=} k_{\text{minor}}, \tag{44}$$

for each $\mu = 1, \ldots, d-1$ for certain constants $f_{\text{minor}} < 1$ and $k_{\text{minor}} \in \mathbb{N}$, subject to the theoretical bound $r_\mu \leq \min(n_\mu r_{\mu-1}, n_{\mu+1} r_{\mu+1})$. Furthermore, a common, upper limit $r_\mu \leq r_{\text{lim}}$ is applied, which is chosen large enough, but likewise in order to avoid unnecessary computation time. The factor f_{minor} is related to the filter in Corollary 2 and the analysis in Sect. 6, but has ultimately been chosen empirically. Whenever necessary, then the rank r_μ is decreased through a simple truncation $\sigma_{r_\mu}^{(\mu)} \leftarrow 0$, while it is increased using a minor singular value $0 < \sigma_{r_\mu+1}^{(\mu)} \ll \sigma_{\min}$. In the latter case, the required, corresponding singular vectors can for example be chosen randomly.

As in the matrix case, the lower limit $\sigma_{\min}$ is a fraction $f_{\sigma_{\min}} \ll 1$ of the residual on the sampling set (cf. Algorithm 4). The parameter $\omega > 0$ is reduced by a factor f_ω in each iteration, slowly reducing the magnitude of regularization. The factor $1 < f_\omega \in (f_\omega^{(\min)}, f_\omega^{(\max)})$, in turn, is increased or decreased after each iteration through a simple heuristic, in such a way that

$$\max_{X \in P, P_2} \frac{\|A^{(\mathtt{iter})} - M\|_X}{\|A^{(\mathtt{iter}-1)} - M\|_X} \stackrel{!}{=} 1 + \varepsilon_{\text{progr}}, \quad \varepsilon_{\text{progr}} > 0, \tag{45}$$

or rather, that it stays close to this fixed value. The tensor $A^{(\mathtt{iter})} = \tau_{r(\mathtt{iter})}(G^{(\mathtt{iter})})$ is the iterate at iteration number `iter`. This adaption ensures that f_ω is not too large as to impair the approximation quality, but neither so small that the required runtime becomes unreasonable.

The algorithm will terminate if one of the following stopping criteria is fulfilled:

- stagnation: $\omega \ll \sigma_{\min}$ and $f_\omega = f_\omega^{(\max)}$
- convergence: $\omega \to 0$ or $\|A^{(\mathtt{iter})} - M\|_P / \|M\|_P \to 0$
- early stop: $\|A^{(\mathtt{iter})} - M\|_{P_2} \gg \min_{i<\mathtt{iter}} \|A^{(i)} - M\|_{P_2}$

We have neglected minor implementation details and practical tweaks in this subsection to focus on the essence of the above criteria, such that we refer to the Matlab code for remaining parts.

8.1 The SALSA algorithm

SALSA (Stable ALS Approximation) for tensors is summarized in Algorithm 4. For the (recommended) choices of tuning parameters also used in the numerical tests, see Sect. 9.4. The Matlab implementation as well as a video showing the rank adaption by means of plotting the singular values during a runtime can be found on the personal webpage of the author Sebastian Krämer, along with all sources that were used to create the presented results.[5] The notion of stability we address in this paper does however not mean a low sensibility to roundoff errors accumulated over multiple iterations. We encountered that even different processor architectures or rearrangement of brackets with regard to associativity in products can change the intermediate approximations. As in almost all cases the algorithm does not find the global minimum, this also holds for the final output, both in case of ALS and SALSA. For once, this is not an actual drawback, since even exact arithmetic would not consistently cause the algorithm to find better local minima, but it should be kept in mind when reconstructing results. The order of computational complexity does not exceed $\mathcal{O}(dr^2|P|)$ using coarse cg, where $r = \max_\mu r_\mu$. Note that the computational complexity per sweep can actually be lower, since not all ranks are kept equal, but some are lower than others.

9 Numerical experiments

We consider the following three algorithms:

- **ALS** (modified Algorithm 3 for $\omega \equiv 0$)
- **SALSA** (Algorithm 4, the algorithm proposed in this work)
- **RTTC** (Riemannian cg for tensor train completion [40])

We explain how ranks are adapted for ALS in Sect. 9.1, shortly present the idea behind RTTC in Sect. 9.2, give details for data acquisition and measurements in Sect. 9.3 as well as tuning parameters in Sect. 9.4. We analyze the results in the latter Sect. 9.9. For

[5] By the time the paper is written, the address is www.igpm.rwth-aachen.de/team/kraemer.

Algorithm 4 SALSA Algorithm

Require: $P \subset \mathscr{I}$, $M|_P$
initialize G s.t. $\tau_r(G) \equiv \|M|_P\|_1/|P|$ for $r \equiv 1$ and $\omega = 1/2\|\tau_r(G)\|_F$
split off a small control set $P_2 \subset P$ (Definition 5)
for iter $= 1, 2, \ldots$ **do**
  proceed SALSA sweep* (Algorithm 3)
  *: and renew lower limit $\sigma_{\min} := f_{\sigma_{\min}} \cdot \frac{|\mathscr{I}|}{|P|} \|\tau_r(G) - M\|_P$
  *: adapt f_ω according to progress (cf. (45))
  *: adapt and decrease ω by factor of f_ω
  adapt rank according to (44) (start this when the first few iteration have passed)
  if a stopping criterion applies (Section 8) **then**
    terminate algorithm
    return iterate for which $\|\tau_r(G) - M\|_{P_2}$ was lowest
  end if
end for

each test, we give a (too large) upper bound r_{lim} for the maximal rank of the iterates, in order to rule out excessive computation times (although this bound is seldomly reached). We would like to emphasize that, in contrast to rank adaption itself, the one dimensional problem of choosing such a bound is easily controlled for example based on the validation set. For simplicity, we use a common mode size $n = n_1 = \cdots = n_d$.

9.1 Rank adaption for standard ALS

Since ALS itself is not rank adaptive, the (so far) most promising approach, that is greedy rank adaption, is chosen. When the progress stagnates, the algorithm searches for the highest (new) singular value $\sigma_+^{(\mu)}$ which any of the rank increases may yield. These values are estimated as follows. Let μ be fixed and G be a representation for which $G^{<\mu-1}$ is column-orthogonal and $G^{>\mu}$ is row-orthogonal. Further, let

$$T := \left(G^{<\mu-1}\right)^T \left((M - \tau_r(G))|_P\right)_{(\mu-1,\mu)} \left(G^{>\mu}\right)^T,$$

$$\alpha_{i_{\mu-1},i_\mu} = \underset{\widetilde{\alpha}_{i_{\mu-1},i_\mu}}{\operatorname{argmin}} \|G^{<\mu-1} \left(G_{\mu-1}(i_{\mu-1} \cdot G_\mu(i_\mu) + \widetilde{\alpha}_{i_{\mu-1},i_\mu} T\left(i_{\mu-1}, i_\mu\right)\right) G^{>\mu} - M_{(\mu-1,\mu)}\|_{P_{(\mu-1,\mu)}(i_{\mu-1},i_\mu)}.$$

We define the core $H(\cdot,\cdot)$, $H(i_{\mu-1}, i_\mu) = \alpha_{i_{\mu-1},i_\mu} T(i_{\mu-1}, i_\mu) \in \mathbb{R}^{r_{\mu-2} \times r_\mu}$ and stack its entries to form the matrix $\mathfrak{H} \in \mathbb{R}^{r_{\mu-2} n_{\mu-1} \times r_\mu n_\mu}$. This yields the candidate $\sigma_+^{(\mu)} := \|\mathfrak{H}\|_2$, the largest singular value of $\mathfrak{H}$. This approach is very similar to the two-fold microsteps as defined in [22] and the rank adaption in AMEn [10], which are both based on DMRG. It however prevents overfitting, since it is equivalent to performing only one, preconditioned CG step (similarly to a Landweber iteration). Furthermore, our experiments suggest that it is more reliable. The corresponding rank $\mu = \operatorname{argmin}_{\widetilde{\mu}} \sigma_+^{(\widetilde{\mu})}$ is increased by 1, using a rank 1 approximation of $\mathfrak{H}$. Since ALS works differently than SALSA, only some stopping criteria can be overtaken, while additional ones are introduced in order to prevent premature termination but also

to avoid unnecessary runtime. No rank decreases are proceeded since this involves tremendous difficulties, of which the most important one is the sheer incapability to decide when and which rank actually to decrease.

9.2 The RTTC algorithm

The article [40], in which RTTC is derived and explained in detail, focuses exclusively on tensor completion using the tensor train format as well (but it can likewise be assumed that it is generalizable to other problems). Instead of alternating optimization, RTTC provides a nonlinear conjugate gradient scheme based on Riemannian optimization, which has comparable computational complexity per sweep. Naturally, the problem of rank adaption also poses a challenge in that setting. Therefore, a heuristic rank adaption is introduced (Algorithm 3 in [40]) which successively tests if a single rank increase yields a tolerable change of the residual on the validation set (based on a parameter $\rho \geq 0$). If so, it continues normally with the next test; otherwise, the algorithm priorly resets to the iterate with previous rank.

We observed however that for example the choice $\rho \neq 0$ worked better given the assignments in Sect. 9.6, but choices other than $\rho = 0$ caused the algorithm to not recover a single instance in case of the rank adaption test tensor in Sect. 9.8. We therefore used three different choices $\rho \in \{0, 0.2, 1\}$ (1 is the default in [40]) and granted RTTC, as opposed to SALSA, the advantage to choose the best result (based on the test set) for each of the following problem classes.

Some minor modification to RTTC were necessary in order to provide fair tests. The allowed number of iterations per rank increase test was increased, since 10 turned out to be too few. In exchange, the relative improvement parameter was raised to 10^{-3}, as lower values did not yield improvements. Since RTTC does not have an actual stopping criterion, we stopped whenever the normalized validation residual increased and at the same time was 1000 times higher than the normalized sampling residual, or whenever the current validation residual was much higher than any previously obtained one (as described in Sect. 8). When terminating, each time the iterate with lowest validation residual was chosen (as described in Algorithm 4). These adaptations were done carefully in order to obtain only improvements in the quality of approximation. We did further not compare the required number of iterations of RTTC, and the algorithm was granted as much time as necessary.

9.3 Data acquisition and measurements

Sampling: In order to obtain a sufficient sampling for each slice of M, we generate the set P in a quasi-random way as follows: For each direction $\mu = 1, \dots, d$ and each index $i_\mu \in \mathcal{I}_\mu$ we pick $c_{\mathrm{sf}} \cdot r_P^2$ indices $i_1, \dots, i_{\mu-1}, i_{\mu+1}, \dots, i_d$ at random (uniformly). This gives in total $|P| = c_{\mathrm{sf}} \cdot dnr_P^2$ samples (excluding duplicate samples). The rank r_P is artificial, such that c_{sf} can be interpreted as sampling factor since the number of degrees of freedom of a TT-tensor of common rank r is slightly less than dnr^2.

Testing: As a test set C, we use a set of the same cardinality as P that is generated in the same way. Of course, neither this set nor the values $M|_C$ are known by the algorithm. The residuals are then measured with respect to the chosen iterate which had the lowest validation residual (cf. Algorithm 4).

Order of optimization: Instead of the sweep we gave before ($\mu = 1, \dots, d$) for simplicity, we alternate between two sweeps ($\mu = 1, \dots, h, \quad \mu = d, \dots, h, \quad h = \lfloor d/2 \rfloor$) to enhance symmetry.

Averaging: With $\langle\cdot\rangle_{\text{ar}}$ we denote the arithmetic mean and by $\langle\cdot\rangle_{\text{geo}}$ the geometric mean which we use for logarithmic scales.

9.4 Implementation details and tuning parameters

All tests for ALS and SALSA were done using a (pure) Matlab implementation. This includes the toolbox *multiprod* [9], which allows a reasonably swift evaluation of products between arrays of matrices, $H(i)J(i)$, $i = 1, \dots, k$, and is much faster than a plain loop. In contrast, some subfunctions of RTTC are based on .mex routines.

Instead of solving full problems in each microstep, both ALS and SALSA use coarse cg (cf. Remark 7), for which the tolerance was empirically chosen low enough such that it did not influence the quality of approximation. Note that the cg steps of RTTC are not comparable, since they are performed on low rank manifolds and used to update all cores at once.

We only list time performances in the appendix, which should be interpreted carefully, while the iteration numbers may provide a clearer picture due to similar computational complexities. All parameters have been chosen equally for all experiments (except ρ for RTTC) with respect to best results, not speed, and could be relaxed for easier problems (or in practice for first trials) to reduce timing considerably. Straightening the tolerances for ALS or RTTC, hence allowing more iterations, did however not lead to notable improvements.

The parameter choices (cf. Sect. 8) for SALSA are given by $\varepsilon_{\text{progr}} = 5 \cdot 10^{-3}$, $f_{\text{minor}} = 0.5$, $k_{\text{minor}} = 2$, $f_\omega^{(\min)} = 1 + 5 \cdot 10^{-4}$, $f_\omega^{(\max)} = 1.1$, $f_{\sigma_{\min}} = 0.1$. The size of the validation set is $|P_2|/(|P| + |P_2|) = 1/20$. These have in parts been chosen empirically and are recommendable for other problems. We observed that any reasonably close values work as well, the more so for larger sampling sets. The performance is in that sense not based on how close the parameters are to some unknown optimal choices. We also refer to the implementation for all details.

9.5 Approximation of a tensor with near uniform singular spectrum

At first, we consider the completion of the following tensor:

$$D(i_1, \dots, i_d) := \left(1 + \sum_{\mu=1}^{d-1} \frac{i_\mu}{i_{\mu+1}}\right)^{-1}, \quad i_\mu = 1, \dots, n, \ \mu = 1, \dots, d$$

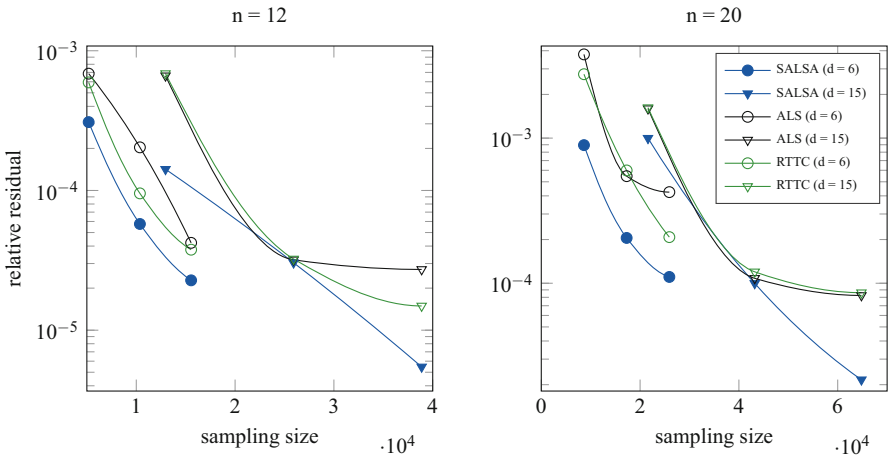

Fig. 5 ($d = 6, 15$, $r_P = 6$, $r_{\text{lim}} = 14$, $n = 12, 20$, $c_{\text{sf}} = 2, 4, 6$) Plotted are, for the tensor D, for varying dimension and mode size, the averaged relative residuals $\langle \|A - M\|_C / \|M\|_C \rangle_{\text{geo}}$ as functions of the sampling size $|P|$ as result of each 20 trials, for ALS (black), SALSA (blue, filled symbols) and RTTC using $\rho = 1$ (green). The markers are exact; the intermediate lines are *shape-preserving piecewise cubic Hermite interpolations of such* (color figure online)

This tensor is not low rank, but has well ordered modes and uniformly exponentially decaying singular values. It can therefore very well be approximated with uniform ranks (for a black box, rank adaptive algorithm however, this is not trivial to recognize) and the low variance of results suggests that mostly a near best approximation is found. Hence, standard ALS can barely be outperformed. The results are plotted in Fig. 5 (See Appendix for Table 1).

9.6 Approximation of three generic tensors with non uniform singular spectrum

We want to demonstrate how different results can be through proper rank adaption, considering the following three tensors, generated by generic functions:

$$f^{(1)}(i_1, \dots, i_8) := \frac{i_1}{4} \cos(i_3 - i_8) + \frac{{i_2}^2}{i_1 + i_6 + i_7} + {i_5}^3 \sin(i_6 + i_3)$$

$$f^{(2)}(i_1, \dots, i_7) := \left(\frac{i_4}{i_2 + i_6} + i_1 + i_3 - i_5 - i_7 \right)^2, \quad i_\mu = 1, \dots, n, \ \mu = 1, \dots, d$$

$$f^{(3)}(i_1, \dots, i_{11}) := \sqrt{i_2 + i_3 + \frac{1}{10}(i_4 + i_5 + i_7 + i_8 + i_9) + \frac{1}{20}(i_1 - i_6 - i_{10} + i_{11})^2};$$

In contrast to the tensor in Sect. 9.5, the modes are not (and hardly can be) ordered in accordance with the TT format. A different ordering may of course yield other results, but we cannot assume to find a better ordering if the approximation fails in the general case. The results are plotted in Fig. 6 (see Appendix for Table 2).

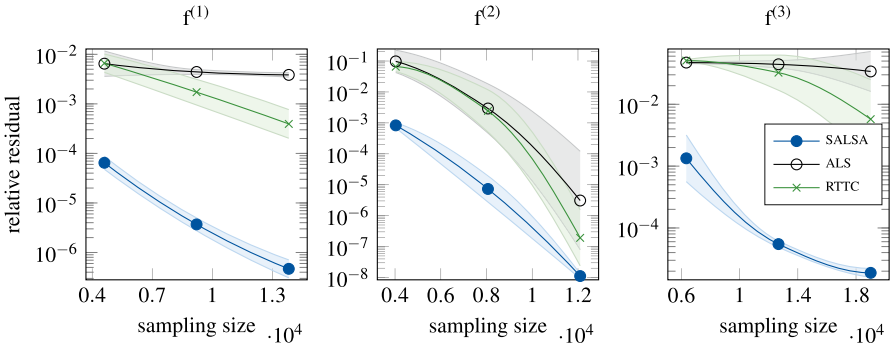

Fig. 6 ($d_1 = 8, d_2 = 7, d_3 = 11, r_P = 6, r_{\lim} = 10, n = 8, c_{\mathrm{sf}} = 2, 4, 6$) Plotted are, for the tensors $f^{(1)}$ (left), $f^{(2)}$ (middle) and $f^{(3)}$ (right), the averaged relative residuals $\langle \|A - M\|_C / \|M\|_C \rangle_{\mathrm{geo}}$ and shadings proportional to the standard deviations as functions of the sampling size $|P|$ as result of each 20 trials, for ALS (black), SALSA (blue, filled symbols) and RTTC using $\rho = 0.2$ (green). The markers are exact; the intermediate lines are *shape-preserving piecewise cubic Hermite interpolations of such* (color figure online)

9.7 Recovery of random tensors with exact low rank

We next consider the recovery of quasi-random tensors with exact low ranks. Although this in practice will never occur, it is a very neutral test.[6] The ranks are generated randomly, but it is ensured that $\langle r \rangle_{\mathrm{ar}} \geq 2/3k$ and $\max(r) \leq k$ for some bound $k \in \mathbb{N}$.

Each of these is generated via a TT representation $A = \tau_r(G)$ where we assign to each entry of each block $G_1, \ldots, G_d$ a uniformly distributed random value in $[-0.5, 0.5]$. Subsequently, the singular values $\Sigma^{(1)}, \ldots, \Sigma^{(d-1)}$ are forced to take uniformly distributed random values in $[0, 1]$ (up to scaling). This is achieved by successive replacements of the current values in G.

As results, we plot the number of successful recoveries ($\|A - M\|_C / \|M\|_C < 10^{-5}$) for different mode sizes n (each single tuple uniform), dimensions d and maximal ranks k of the target tensor (Figs. 7 and 8).

9.8 Recovery of the rank adaption test tensor

Last but not least, we consider the recovery of tensors as in Example 3, for which Q_1, Q_4, Q_5 and Q_6 are generated quasi-randomly for each trial. For an explanation of the results in Fig. 9, we refer to Sect. 9.7.

9.9 Analysis of results

SALSA is superior in nearly all observed cases. For tensors which could as well be approximated with uniform ranks, the differences are marginal as is to be expected.

[6] Note that in some papers, uniform distributions on $[0, 1]$ are used such that all entries of the target tensor are positive, causing each first singular value to be huge compared to all following ones. This leads to a tremendous simplification of the completion problem. There seems to be no indication yet that the sampling required for the completion of a random tensor is in general close to what is stated for the matrix case [8].

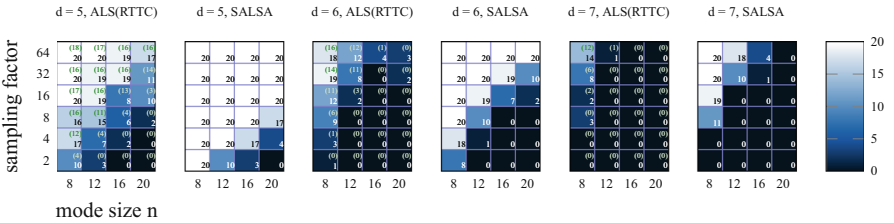

Fig. 7 ($d = 5, 6, 7, r_P = 6, r_{\text{lim}} = 9, n = 8, 12, 16, 20, c_{\text{sf}} = 2, 4, 8, 16, 32, 64$) Displayed as 20 shades of blue (black (0) to white (all 20)) are the number of successful reconstructions for random tensors with maximal rank $k = 6$ for ALS and SALSA. The bracketed green numbers are the results for RTTC using $\rho = 0.2$ and are independent of the shading. We recommend use of the digital version for better readability (color figure online)

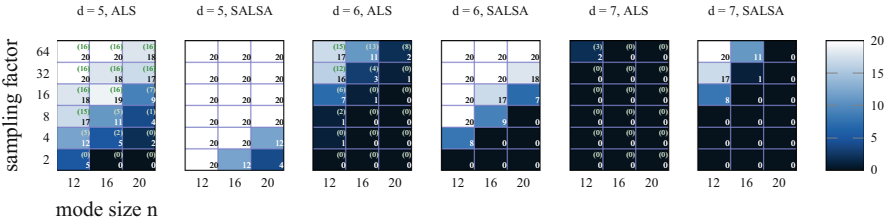

Fig. 8 ($d = 5, 6, 7, r_P = 8, r_{\text{lim}} = 11, n = 12, 16, 20, c_{\text{sf}} = 2, 4, 8, 16, 32, 64$) Displayed as 20 shades of blue (black (0) to white (all 20)) are the number of successful reconstructions for random tensors with maximal rank $k = 8$ for ALS and SALSA. The bracketed green numbers are the results for RTTC using $\rho = 0.2$ and are independent of the shading. We recommend use of the digital version for better readability (color figure online)

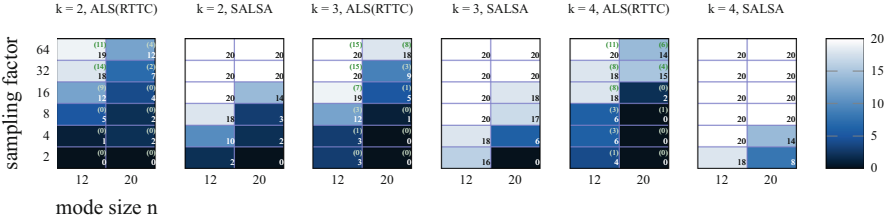

Fig. 9 ($d = 6, r_P = 2k, r_{\text{lim}} = 2k + 3, n = 12, 20, c_{\text{sf}} = 2, 4, 8, 16, 32, 64$) Displayed as 20 shades of blue (black (0) to white (all 20)) are the number of successful reconstructions for the rank adaption test tensor with rank $(1, k, k, k, 1, 2k, 1)$ for ALS and SALSA. The bracketed green numbers are the results for RTTC using $\rho = 0$ and are independent of the shading. We recommend use of the digital version for better readability (color figure online)

The three generic functions show that the residuals can be multiple orders of magnitude better, and although the functions were chosen quite randomly, we do of course not want to over-interpret these specific results. Finally, for the more neutral test of random tensor recovery, the required sampling seems to be overall 4 to 8 times lower. For the rank adaption test tensor, the performance of SALSA becomes even better for larger rank k (this is due to the larger total sampling), while greedy ALS runs into the predicted trouble, as does RTTC. In general, RTTC performs slightly better than greedy ALS in the approximation of tensors with exponentially declining singular values, while

the latter is slightly better in the random recovery tests. This difference is unlikely because of their optimization techniques, which are known to achieve similar results, but rather due to the different rank adaption heuristics. This insight is also based on closer inspections of single tests, which suggest that the rank adaption of RTTC is inferior even to the one of ALS in the majority of cases.

10 Conclusions

In this article, we have demonstrated that the most successful completion algorithms are very sensitive to rank changes and that existing rank adaption methods suffer from this.
In order to correct this, as proven for SALSA, we suggested a regularization motivated by averaged microsteps in order to uncouple the optimization of a discrete, technical rank. While there is likely room for improvements and rigorous convergence bounds remain subject to future work, we take the noteworthy numerical results as indication that *stability (under truncation)* is a worthwhile property. The computational complexity of SALSA is further reduced to the minimal order through use of a coarse cg method. Although we focused on tensor completion (with possibly small sampling sets), the derivations given in this paper allow for a generalization to other semi-elliptic problems. Furthermore, it may be possible to adapt the presented ideas to manifold based methods such as RTTC.

Acknowledgements We thank the anonymous referees whose comments and suggestions helped to improve and clarify this manuscript.

11 Appendix (Experimental data)

Following are the precise values for Figs. 5 and 6, for $R_C := \|A - M\|_C$ and $R_P := \|A - M\|_P$.
See Tables 1 and 2.

Table 1 Results for Sect. 9.5 (with arithmetic and geometric variances in brackets) using a (pure) Matlab implementation

d	c_{sf}	ALS				SALSA			
		$\langle R_C/\|M_C\|\rangle_{\mathrm{geo}}$	$\langle R_P/\|M_P\|\rangle_{\mathrm{geo}}$	$\langle\mathrm{time}\rangle_{\mathrm{ar}}$	$\langle\mathrm{iter}\rangle_{\mathrm{ar}}$	$\langle R_C/\|M_C\|\rangle_{\mathrm{geo}}$	$\langle R_P/\|M_P\|\rangle_{\mathrm{geo}}$	$\langle\mathrm{time}\rangle_{\mathrm{ar}}$	$\langle\mathrm{iter}\rangle_{\mathrm{ar}}$
$n = 12$									
6	2	6.9e−04 (2.1)	1.1e−04 (2.9)	85 (57)	467 (151)	3.1e−04 (1.1)	6.8e−06 (3.7)	81 (10)	412 (33)
	4	2.0e−04 (1.9)	3.9e−05 (3.8)	131 (86)	527 (154)	5.8e−05 (1.5)	1.4e−06 (2.0)	170 (29)	546 (58)
	6	4.2e−05 (1.8)	4.7e−06 (2.8)	215 (76)	624 (108)	2.3e−05 (1.4)	1.5e−06 (1.4)	209 (23)	581 (41)
9	2	7.6e−04 (2.1)	2.3e−04 (4.4)	145 (88)	606 (198)	1.8e−04 (1.1)	6.9e−06 (2.8)	129 (9)	455 (23)
	4	8.0e−05 (1.1)	2.0e−05 (1.1)	276 (57)	804 (81)	2.7e−05 (1.2)	7.4e−07 (2.0)	298 (34)	644 (40)
	6	6.6e−05 (1.2)	2.5e−05 (1.2)	355 (94)	834 (105)	9.6e−06 (1.2)	5.5e−07 (1.5)	457 (50)	732 (48)
15	2	6.6e−04 (1.5)	3.3e−04 (2.1)	384 (113)	951 (159)	1.4e−04 (1.4)	1.1e−05 (1.8)	266 (20)	498 (31)
	4	3.2e−05 (1.1)	6.8e−06 (1.1)	961 (62)	1396 (44)	3.0e−05 (1.1)	4.3e−06 (1.6)	540 (35)	642 (32)
	6	2.7e−05 (1.1)	8.3e−06 (1.1)	1278 (85)	1416 (47)	5.5e−06 (1.8)	2.9e−07 (1.6)	1133 (92)	827 (48)
$n = 20$									
6	2	3.8e−03 (1.1)	1.7e−03 (1.1)	89 (38)	378 (93)	8.9e−04 (1.3)	3.4e−05 (2.9)	119 (15)	375 (34)
	4	5.5e−04 (1.2)	1.7e−04 (1.2)	157 (45)	468 (74)	2.1e−04 (1.3)	4.0e−06 (1.5)	256 (30)	506 (36)
	6	4.2e−04 (1.2)	1.6e−04 (1.4)	198 (62)	483 (80)	1.1e−04 (1.4)	5.0e−06 (1.3)	329 (27)	536 (32)
9	2	2.6e−03 (1.0)	1.4e−03 (1.1)	204 (101)	536 (141)	4.6e−04 (1.1)	2.4e−05 (2.2)	224 (15)	452 (26)
	4	2.9e−04 (1.1)	1.0e−04 (1.1)	426 (71)	758 (73)	1.3e−04 (1.3)	2.7e−06 (1.5)	452 (57)	583 (38)
	6	2.0e−04 (1.1)	7.2e−05 (1.4)	618 (151)	818 (103)	4.2e−05 (1.1)	2.0e−06 (1.3)	748 (53)	692 (23)
15	2	1.6e−03 (1.0)	8.6e−04 (1.0)	698 (208)	960 (166)	1.0e−03 (1.9)	1.8e−04 (6.0)	361 (64)	417 (46)
	4	1.1e−04 (1.1)	2.3e−05 (1.2)	1766 (123)	1402 (53)	1.0e−04 (1.1)	1.5e−05 (1.5)	899 (50)	608 (27)
	6	8.2e−05 (1.0)	2.8e−05 (1.1)	2299 (139)	1392 (45)	2.2e−05 (1.2)	9.2e−07 (1.4)	1873 (188)	782 (44)

For both ALS and SALSA, coarse CG is used. Note that most iterations are performed while the rank is not at its maximum yet

Table 2 Results for Sect. 9.6 (with arithmetic and geometric variances in brackets) using a (pure) Matlab implementation

$n = 12$		ALS				SALSA			
d	c_{sf}	$\langle R_C/\|M_C\|\rangle_{\mathrm{geo}}$	$\langle R_P/\|M_P\|\rangle_{\mathrm{geo}}$	$\langle\mathrm{time}\rangle_{\mathrm{ar}}$	$\langle\mathrm{iter}\rangle_{\mathrm{ar}}$	$\langle R_C/\|M_C\|\rangle_{\mathrm{geo}}$	$\langle R_P/\|M_P\|\rangle_{\mathrm{geo}}$	$\langle\mathrm{time}\rangle_{\mathrm{ar}}$	$\langle\mathrm{iter}\rangle_{\mathrm{ar}}$
8	2	6.5e−03 (3.2)	5.3e−03 (3.4)	72 (46)	423 (139)	6.5e−05 (1.8)	1.1e−06 (5.8)	69 (16)	359 (66)
	4	4.4e−03 (1.2)	4.1e−03 (1.3)	34 (34)	258 (114)	3.7e−06 (1.8)	1.9e−08 (5.5)	133 (32)	497 (87)
	6	3.8e−03 (1.2)	3.6e−03 (1.2)	47 (40)	288 (122)	4.7e−07 (2.3)	4.5e−09 (4.0)	157 (34)	520 (88)
7	2	9.7e−02 (5.2)	4.0e−02 (6.5)	50 (54)	322 (167)	8.3e−04 (1.5)	9.7e−06 (3.1)	65 (11)	345 (37)
	4	2.9e−03 (38.7)	4.7e−04 (151.0)	104 (102)	482 (260)	7.3e−06 (7.0)	1.3e−07 (6.7)	130 (40)	506 (117)
	6	3.1e−06 (1506.0)	1.2e−06 (2386.6)	125 (79)	576 (271)	1.1e−08 (1.4)	3.6e−09 (1.0)	175 (17)	579 (48)
11	2	4.7e−02 (1.1)	4.6e−02 (1.1)	66 (72)	291 (166)	1.3e−03 (5.7)	1.9e−04 (14.1)	53 (16)	185 (55)
	4	4.4e−02 (1.3)	4.1e−02 (1.5)	74 (76)	288 (166)	5.5e−05 (1.3)	2.9e−06 (2.4)	138 (19)	328 (34)
	6	3.4e−02 (4.4)	3.0e−02 (6.4)	115 (167)	319 (256)	1.9e−05 (1.4)	7.2e−07 (2.5)	248 (50)	408 (56)

For both ALS and SALSA, coarse CG is used. Note that most iterations are performed while the rank is not at its maximum yet

References

1. Bachmayr, M., Schneider, R.: Iterative methods based on soft thresholding of hierarchical tensors. Found. Comput. Math. (2016). https://doi.org/10.1007/s10208-016-9314-z
2. Bachmayr, M., Schneider, R., Uschmajew, A.: Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations. Found. Comput. Math. (2016). https://doi.org/10.1007/s10208-016-9317-9
3. Ballani, J., Grasedyck, L.: A projection method to solve linear systems in tensor format. Numer. Linear Algebra Appl. **20**(1), 27–43 (2013). https://doi.org/10.1002/nla.1818
4. Ballani, J., Grasedyck, L., Kluge, M.: Black box approximation of tensors in hierarchical tucker format. Linear Algebra Appl. **438**(2), 639–657 (2013). https://doi.org/10.1016/j.laa.2011.08.010
5. Beylkin, G., Mohlenkamp, M.: Numerical operator calculus in higher dimensions. PNAS **99**(16), 10246–10251 (2002). https://doi.org/10.1073/pnas.112329799
6. Burke, J.V., Lewis, A.S., Overton, M.L.: A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. SIAM J. Optim. **15**(3), 751–779 (2005). https://doi.org/10.1137/030601296
7. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717 (2009). https://doi.org/10.1007/s10208-009-9045-5
8. Candès, E.J., Tao, T.: The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theor. **56**(5), 2053–2080 (2010). https://doi.org/10.1109/TIT.2010.2044061
9. de Leva, P.: multiprod—multiple matrix multiplications, with array expansion enabled (2010). https://www.mathworks.com/matlabcentral/fileexchange/8773-multiple-matrix-multiplications-with-array-expansion-enabled
10. Dolgov, S.V., Savostyanov, D.V.: Alternating minimal energy methods for linear systems in higher dimensions. SIAM J. Sci. Comput. **36**(5), A2248–A2271 (2014). https://doi.org/10.1137/140953289
11. Dopico, F.M.: A note on sin θ theorems for singular subspace variations. BIT Numer. Math. **40**(2), 395–403 (2000). https://doi.org/10.1023/A:1022303426500
12. Espig, M., Khachatryan, A.: Convergence of alternating least squares optimisation for rank-one approximation to high order tensors (2015). arXiv:1503.05431
13. Gandy, S., Recht, B., Yamada, I.: Tensor completion and low-n-rank tensor recovery via convex optimization. Inverse Problems **27**(2), 025010 (2011)
14. Grasedyck, L.: Hierarchical singular value decomposition of tensors. SIAM J. Matrix Anal. Appl. **31**(4), 2029–2054 (2010). https://doi.org/10.1137/090764189
15. Grasedyck, L., Kluge, M., Krämer, S.: Variants of alternating least squares tensor completion in the tensor train format. SIAM J. Sci. Comput. **37**(5), A2424–A2450 (2015). https://doi.org/10.1137/130942401
16. Grasedyck, L., Kressner, D., Tobler, C.: A literature survey of low-rank tensor approximation techniques. GAMM-Mitteilungen **36**(1), 53–78 (2013). https://doi.org/10.1002/gamm.201310004
17. Gross, D.: Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**(3), 1548–1566 (2011). https://doi.org/10.1109/TIT.2011.2104999
18. Hackbusch, W.: Numerical tensor calculus. Acta Numer. **23**, 651–742 (2014). https://doi.org/10.1017/S0962492914000087
19. Hackbusch, W., Kühn, S.: A new scheme for the tensor representation. J. Fourier Anal. Appl. **15**(5), 706–722 (2009). https://doi.org/10.1007/s00041-009-9094-9
20. Hackbusch, W., Schneider, R.: Tensor Spaces and Hierarchical Tensor Representations, pp. 237–261. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-08159-5_12
21. Hastie, T., Mazumder, R., Lee, J.D., Zadeh, R.: Matrix completion and low-rank svd via fast alternating least squares. J. Mach. Learn. Res. **16**(1), 3367–3402 (2015)
22. Holtz, S., Rohwedder, T., Schneider, R.: The alternating linear scheme for tensor optimization in the tensor train format. SIAM J. Sci. Comput. **34**(2), A683–A713 (2012). https://doi.org/10.1137/100818893
23. Holtz, S., Rohwedder, T., Schneider, R.: On manifolds of tensors of fixed TT-rank. Numer. Math. **120**(4), 701–731 (2012). https://doi.org/10.1007/s00211-011-0419-7
24. Jain, P., Netrapalli, P., Sanghavi, S.: Low-rank matrix completion using alternating minimization. In: Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing, STOC '13, pp. 665–674. ACM, New York, NY, USA (2013). https://doi.org/10.1145/2488608.2488693
25. Jeckelmann, E.: Dynamical density-matrix renormalization-group method. Phys. Rev. B **66**, 045,114 (2002). https://doi.org/10.1103/PhysRevB.66.045114

26. Krämer, S.: The geometrical description of feasible singular values in the tensor train format (2017). arXiv:1701.08437
27. Kressner, D., Steinlechner, M., Vandereycken, B.: Low-rank tensor completion by riemannian optimization. BIT Numer. Math. **54**(2), 447–468 (2014). https://doi.org/10.1007/s10543-013-0455-z
28. Liu, Y., Shang, F.: An efficient matrix factorization method for tensor completion. IEEE Signal Process. Lett. **20**(4), 307–310 (2013). https://doi.org/10.1109/LSP.2013.2245416
29. Matthies, H.G., Zander, E.: Solving stochastic systems with low-rank tensor compression. Linear Algebra Appl. **436**(10), 3819–3838 (2012). https://doi.org/10.1016/j.laa.2011.04.017
30. Mirsky, L.: Symmetric gauge functions and unitarily invariant norms. Q. J. Math. **11**(1), 50–59 (1960). https://doi.org/10.1093/qmath/11.1.50
31. Mu, C., Huang, B., Wright, J., Goldfarb, D.: Square deal: lower bounds and improved relaxations for tensor recovery. In: Jebara, T., Xing, E.P., (eds.) Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 73–81. JMLR Workshop and Conference Proceedings (2014). http://jmlr.org/proceedings/papers/v32/mu14.pdf
32. Oseledets, I., Tyrtyshnikov, E.: TT-cross approximation for multidimensional arrays. Linear Algebra Appl. **432**(1), 70–88 (2010). https://doi.org/10.1016/j.laa.2009.07.024
33. Oseledets, I.V.: Tensor-train decomposition. SIAM J. Sci. Comput. **33**(5), 2295–2317 (2011). https://doi.org/10.1137/090752286
34. Oseledets, I.V., Tyrtyshnikov, E.E.: Breaking the curse of dimensionality, or how to use svd in many dimensions. SIAM J. Sci. Comput. **31**(5), 3744–3759 (2009). https://doi.org/10.1137/090748330
35. Rauhut, H., Schneider, R., Stojanac, Ž.: Tensor Completion in Hierarchical Tensor Representations, pp. 419–450. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16042-9_14
36. Recht, B.: A simpler approach to matrix completion. J. Mach. Learn. Res. **12**, 3413–3430 (2011)
37. Rohwedder, T., Uschmajew, A.: On local convergence of alternating schemes for optimization of convex problems in the tensor train format. SIAM J. Numer. Anal. **51**(2), 1134–1162 (2013). https://doi.org/10.1137/110857520
38. Signoretto, M., TranDinh, Q., De Lathauwer, L., Suykens, J.A.K.: Learning with tensors: a framework based on convex optimization and spectral regularization. Mach. Learn. **94**(3), 303–351 (2014). https://doi.org/10.1007/s10994-013-5366-3
39. Silva, C.D., Herrmann, F.J.: Optimization on the hierarchical tucker manifold applications to tensor completion. Linear Algebra Appl. **481**, 131–173 (2015). https://doi.org/10.1016/j.laa.2015.04.015
40. Steinlechner, M.: Riemannian optimization for high-dimensional tensor completion. SIAM J. Sci. Comput. **38**(5), S461–S484 (2016). https://doi.org/10.1137/15M1010506
41. Vidal, G.: Efficient classical simulation of slightly entangled quantum computations. Phys. Rev. Lett. **91**, 147,902 (2003). https://doi.org/10.1103/PhysRevLett.91.147902
42. Wedin, P.Å.: Perturbation bounds in connection with singular value decomposition. BIT Numer. Math. **12**(1), 99–111 (1972). https://doi.org/10.1007/BF01932678
43. Wen, Z., Yin, W., Zhang, Y.: Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. Math. Program. Comput. **4**(4), 333–361 (2012). https://doi.org/10.1007/s12532-012-0044-1
44. Weyl, H.: Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). Math. Ann. **71**(4), 441–479 (1912). https://doi.org/10.1007/BF01456804
45. White, S.R.: Density matrix formulation for quantum renormalization groups. Phys. Rev. Lett. **69**, 2863–2866 (1992). https://doi.org/10.1103/PhysRevLett.69.2863