PREPRINT 461

JANUARY 2017

Stable ALS Approximation in the TT-Format for Rank-Adaptive Tensor Completion

Lars Grasedyck* and Sebastian Krämer*

Institut für Geometrie und Praktische Mathematik
Templergraben 55, 52062 Aachen, Germany

* IGPM, RWTH Aachen University, Templergraben 55, 52056 Aachen, +49-241-80-92317 (fax)
L. Grasedyck: lgr@igpm.rwth-aachen.de: +49-241-80-97069 (tel)
S. Krämer (corresponding author): kraemer@igpm.rwth-aachen.de: +49-241-80-94875 (tel)

Stable ALS Approximation in the TT-Format for Rank-Adaptive Tensor Completion

Lars Grasedyck* Sebastian Krämer*

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 flaw as *instability (under truncation)*. As a consequence of this flaw, arbitrarily small changes in the singular values of an 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 micro-steps. We prove its *stability* and derive a natural semi-implicit rank adaption strategy. We further prove that the standard ALS micro-steps are only stable on manifolds of fixed ranks, and only around points that have what we define as *internal tensor restricted isometry property iTRIP*. Finally, we provide numerical examples that show improvements of the reconstruction quality up to orders of magnitude in the new Stable ALS Approximation (SALSA) compared to standard ALS.

Keywords. tensor completion, MPS, tensor train, TT, hierarchical Tucker, HT, alternating optimization, ALS, high-dimensional, low rank, SVD, ill-posedness, stability

AMS subject classifications. 15A18, 15A69, 65F10, 65F22, 90C06, 90C31

1 Introduction

Data sparse formats for high-dimensional tensors are typically based on notions of (low) rank(s) and non-unique representations with correspondingly few degrees of freedom - for comprehensive survey articles, we refer to [14, 16, 18]. These representations can be understood as technical tools to generate tensors, which are the main, or sole, objects of interest. A method may be based on these representations, such as ALS, and follow the same concept for any fixed rank. However, through the correspondence of data and full tensor, the method may also yield an accordant map acting on the full tensor space.
In the setting of matrix completion, the data model, or tensor format $\mathcal{T}$, often is the low rank representation, i.e. a function $\tau_r : (X, Y) \mapsto XY^T \in \mathbb{R}^{n\times m}$ for $(X, Y) \in \mathcal{D}_r :=$

*IGPM, RWTH Aachen University, Templergraben 55, 52056 Aachen, +49-241-80-92317 (fax)
L. Grasedyck: lgr@igpm.rwth-aachen.de: +49-241-80-97069 (tel)
S. Krämer (corresponding author): kraemer@igpm.rwth-aachen.de: +49-241-80-94875 (tel)

$\mathbb{R}^{n\times r}\times\mathbb{R}^{m\times r}$. A method $\mathcal{M}$ applied to this model may be the least squares optimization of Y, i.e.

$$\mathcal{M}_r(X,Y)=(X,\operatorname*{argmin}_{\widetilde{Y}}\|X\widetilde{Y}^T-M\|_P), \tag{1.1}$$

where P is the sampling set. In order to obtain a well defined function, the minimization of $\|X\widetilde{Y}^T\|_F$ serves as secondary criterion if there is not yet a unique argument of the minimum. Formally, for each value of the classical matrix rank r, these are different functions. Our starting point as well as the central objective of this paper is the continuity of these algorithmic steps as functions on the whole tensor space

$$A\in\mathbb{R}^{\mathcal{I}},\quad \mathcal{I}:=\mathcal{I}_1\times\cdots\times\mathcal{I}_d,\quad \mathcal{I}_\mu:=\{1,\dots,n_\mu\},\quad \mu\in D:=\{1,\dots,d\}.$$

Definition 1.1 (Stability)**.** *Let $\mathcal{T}$ be a tensor format in respect of which every tensor has a unique rank $r\in\mathbb{N}_0^m$ and hence belongs to one of the disjoint subsets $\mathcal{T}(r)\subset\mathbb{R}^{\mathcal{I}}$. Let further $\tau_r:\mathcal{D}_r\to\mathbb{R}^{\mathcal{I}}$ be the function that maps a representation to its tensor and $\mathcal{M}$ be a method that maps any rank r to a function $\mathcal{M}_r:\mathcal{D}_r\to\mathcal{D}_r$ (the optimization method for fixed rank). We define the following properties:*

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

$$f_{\mathcal{M}}:\mathbb{R}^{\mathcal{I}}\to\mathbb{R}^{\mathcal{I}},\quad f_{\mathcal{M}}(A):=\tau_{r(A)}\circ\mathcal{M}_{r(A)}\circ\tau_{r(A)}^{-1}(A), \tag{1.2}$$

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

Certainly, stability implies any fix-rank stability. Properly calibrating the rank r for unstable methods poses a very intricate problem. Most of the operators applied to representations are stable, e.g. truncations based on matrix singular values. The situation changes if we apply a partial optimization (or micro-) step on a low rank representation, which is a very common step for many algorithms.

Example 1.2 (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 $A=A(\varepsilon)$*

$$M:=\begin{pmatrix}\boxed{?} & 1.1 & 1\\ 1 & 1 & 1\\ 1.1 & 1 & 1\end{pmatrix},\quad A(\varepsilon):=\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix}+\varepsilon\begin{pmatrix}a&0&-a\\1+a&0&-1-a\\1-a&0&-1+a\end{pmatrix},$$

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 (1.1), we replace $Y(\varepsilon)$ of the low rank representation $A(\varepsilon)=X(\varepsilon)Y(\varepsilon)^T$ by the local minimizer, where in this case

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

This optimization yields a new matrix, $B(\varepsilon) = \tau_r \circ \mathcal{M}_r \circ \tau_r^{-1}(A(\varepsilon))$, given by

$$B(0) = \begin{pmatrix} 1.05 & * & * \\ 1.05 & * & * \\ 1.05 & * & * \end{pmatrix}, \qquad B(\varepsilon) = \begin{pmatrix} 1 + \frac{1}{20a} & * & * \\ 1.0 & * & * \\ 1.1 & * & * \end{pmatrix} \text{ if } \varepsilon > 0. \quad (* \text{ is some value})$$

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. One minor detail is that the rank two approximation $B(1)$ diverges as $a \to 0$, in particular it is not convergent although the initial guess $A(1)$ converges to a rank two matrix as $a \to 0$. Thus, the micro-step 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 $a \to 0$. The unboundedness can be remedied by adding a regularization term in the least squares functional, e.g. $+\|XY^T\|$, but the ALS step remains unstable.

This example likewise demonstrates that ALS for tensor completion is not stable. It is easy to see that this is not an exceptional problem, but occurs systematically (cf. Example 2.1). For the rest of this article we consider the problem of (approximately) reconstructing a tensor from a given data set $M_P = \{M_i\}_{i \in P}$, where $P \subset \mathcal{I}$ is fixed. For the underlying tensor M it is assumed that there exists some (low) rank $r \in \mathbb{N}^{d-1}$ to yield an approximation $M_\varepsilon \in TT(r)$ where $TT(r)$ is the tensor train [28, 29] (or matrix product state (MPS) [35, 39], or special case of the hierarchical [12, 17]) format:

Definition 1.3 (TT format)**.** *A tensor $A \in \mathbb{R}^{\mathcal{I}}$ is in the set $TT(r)$, $r \in \mathbb{N}^{d-1}$ if for $\mu = 1, \dots, d$ and $i_\mu \in \mathcal{I}_\mu$ there exist $G_\mu(i_\mu) \in \mathbb{R}^{r_{\mu-1} \times r_\mu}$ ($r_0 = r_d = 1$) such that*

$$A(i_1, \dots, i_d) = G_1(i_1) \cdots G_d(i_d), \qquad i \in \mathcal{I}.$$

The representation of A in this form is shortly called the TT(r) or TT format. If we want to stress the dependency of A on the so-called cores G_μ then we write $A = \tau_r(G) := G_1 \boxtimes \ldots \boxtimes G_d$, where we define $\boxtimes$ for two cores H_1, H_2 as $(H_1 \boxtimes H_2)(i, j) := H_1(i)\ H_2(j)$ (interpreting TT-cores as vectors of matrices). For the matrix Kronecker product we use the symbol $\otimes$.

We do not use any more information, explicitly **no detailed knowledge about the rank $\mathbf{r} = (\mathbf{r_1}, \dots, \mathbf{r_{d-1}})$ is assumed** and we will demonstrate why this can be troublesome in the following.

Notation 1.4 (TT singular values)**.** *Analogously to the TT-ranks, we define the TT-singular values $\sigma = (\sigma^{(1)}, \dots, \sigma^{(d-1)})$ of a tensor A as the unique singular values of the corresponding matricizations $A^{(1,\dots,\mu)} \in \mathbb{R}^{(\mathcal{I}_1 \times \cdots \times \mathcal{I}_\mu) \times (\mathcal{I}_{\mu+1} \times \cdots \times \mathcal{I}_d)}$ with entries*

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

of A, such that $\sigma^{(\mu)}$ contains the ordered singular values of $A^{(1,\dots,\mu)}$, $\mu = 1, \dots, d-1$. Hence, 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 and σ the (TT-)singular spectrum.

Many tensors of relevance have very well and uniformly behaving singular values, but this is certainly not the general case, as the following example demonstrates. One can even prove that there is no limitation to the shape of the singular spectrum for fixed rank, except the trivial $\|\sigma^{(i)}\|_2^2 = \|\sigma^{(j)}\|_2^2$, $\forall i, j$, provided the mode sizes $n_1, \dots, n_d$ are large enough [22].

Example 1.5 (Rank adaption test tensor)**.** *For $k \in \mathbb{N}$, let $Q \in \mathbb{R}^{n_1 \times \ldots \times n_4}$ be an orthogonally decomposable 4-dimensional TT Tensor with rank (k, k, k) and uniform singular*

values $\sigma^{(1)} = \sigma^{(2)} = \sigma^{(3)} = (\alpha, \alpha, \ldots)$ *as well as* $B \in \mathbb{R}^{n_5 \times n_6}$ *be a rank* $2k$ *matrix with exponentially decaying singular values* $\sigma^{(5)} \propto (\beta^{-1}, \beta^{-2}, \ldots)$ *for some* $\alpha, \beta > 0$. *Then the separable tensor* $A \in \mathbb{R}^{n_1 \times \cdots \times n_6}$ *defined by* $A(i) = Q(i_1, \ldots, i_4) \cdot B(i_5, i_6)$ *has singular values* σ *and rank* $r = (k, k, k, 1, 2k)$. *For an explicit construction, see Appendix A.*

By definition, A 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 ($\mu = 1, \ldots, d-1$) 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 [10], approximation of orthogonally decomposable tensors with lower rank can be problematic. In numerical tests, 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 ($\mu = 1, \ldots, d-1$) starting from $r \equiv 1$ and decrease ranks when the corresponding singular values are below a threshold. The problem with this strategy is quite obvious, namely that $r_5 = 1$. If this rank is overestimated, the observed sampling points will be misinterpreted (oversampling) and it does not matter how small corresponding singular values become (see Lemma 2.1).

These indicated difficulties gain more importance with high dimension, but for one micro-step at a time, can be resolved by regarding only three components of a tensor. We will come back to this in Section 3.

1.1 Relation to Other Tensor Methods

Whenever a tensor is point-wise available, algorithms such as the TT-SVD [28] 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, 27].

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, 25], which naturally increase the rank and rely on subsequent reductions, or also by rank preservative optimization, such as alternating optimization [8, 10, 19, 32], 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 [30]. 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 [19, 21], are misled.

Nuclear norm minimization, being closely related to compressed sensing as well, for the matrix case [6, 7, 15, 31] has a very strong theoretical background, yet the simplifications required for an adaption to tensors [11, 24, 33] do not seem to allow for an appropriate generalization [26]. While these approaches rely on a direct adaption of the target

function, that is convex relaxation, our starting point are the micro steps provided by alternating least squares. In that sense, we treat each update and adaption as part of a learning progress.
For fixed or uniform rank, there have been proposals in hierarchical, or tree-, formats [23, 34] as well, the essential adaption of which however 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 [13, 37] and its regularization properties are a first starting point for this work.

The rest of the article is organized as follows: In Section 2, we further investigate instability and exemplarily analyze approaches towards it in the matrix case. In Section 3, we introduce further notations and thereby reduce the setting to essential three dimensions. In the main Section 4, we continue with the previously carried out ideas and thereby motivate a modified, iterate dependent residual function. We then derive its minimizer and prove stability (Theorem 4.13) for the thereby obtained regularized micro-steps. Subsequently, in Section 5, these results are transferred back to arbitrarily dimensional tensors. Section 6 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 Section 7. Appendix A provides remaining proofs and in Appendix B, we shortly analyze a key element of SALSA. Appendix C includes detailed, experimental data.

2 Instability and Approaches to Resolve the Problem

As previously mentioned, instability poses a systematic flaw in ALS, or for that matter, in any such range based optimization:

Example 2.1 (ALS is unstable)**.** *Consider the micro-step $\mathcal{M}$ as in* (1.1)*. Let U, V be orthogonal, such that $U\Sigma V^T$ is a truncated SVD of a rank r matrix $A = \tau_r(U, V\Sigma^T)$. We now let $\sigma_r \to 0$ such that $A^* := A|_{\sigma_r=0}$ has rank $r-1$. The update is independent of the value $\sigma_r > 0$:*

$$f_{\mathcal{M}}(A) = \tau_r(\mathcal{M}_r(U, V\Sigma^T)) = U \operatorname*{argmin}_Y \|UY^T - M\|_P = \lim_{\varepsilon \searrow 0} f_{\mathcal{M}}(A|_{\sigma_r=\varepsilon}) \tag{2.1}$$

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

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

is in general different from the limit (2.1)*. The same holds for an analogous update of $X = U\Sigma$.*

The micro-steps 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 the core G_μ. Modifying the micro-steps such that stability is gained is one task. Another aspect, however, is that we aim for a natural way to do this, which we will discuss in the following.
A quite successful approach for completion has been the rank increasing strategy, e.g. [37]. By the given limitation to all ranks, a regularization is introduced to the target function. High frequencies, corresponding to low singular values, are excluded up to a certain progress.

A similar kind of effect can be achieved by assuming an uncertainty of 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 *technical* rank that is currently used. We will first view this in a minimal fashion for the matrix case and the method $\mathcal{M}$ defined by (1.1). With this approach, we can motivate an algorithm that is stable under truncation and allows to straightforwardly adapt ranks nonuniformly. It optimizes, in a loose sense, continuously between manifolds of different ranks.
Assuming local integrability of $f_{\mathcal{M}}$ (as defined in (1.2)), we obtain that the averaged function

$$f^*_{\mathcal{M}}(A) := \frac{1}{|\mathbb{V}_{A,\omega}|} \int_{\mathbb{V}_{A,\omega}} f_{\mathcal{M}}(H)\,\mathrm{d}H \tag{2.2}$$
$$\mathbb{V}_{A,\omega} := \{H \in image(\tau_{\widetilde{r}}) \mid \|H - A\| \leq \omega\}$$

is continuous within $image(\tau_{\widetilde{r}})$, where $\widetilde{r}$ may be considered an upper bound to the rank, cf. Figure 1. However, this function does not preserve low rank structure and therefore we cannot find a method $\mathcal{M}^*$ for which $f^*_{\mathcal{M}} = f_{\mathcal{M}^*}$. Consider instead a scenario in which

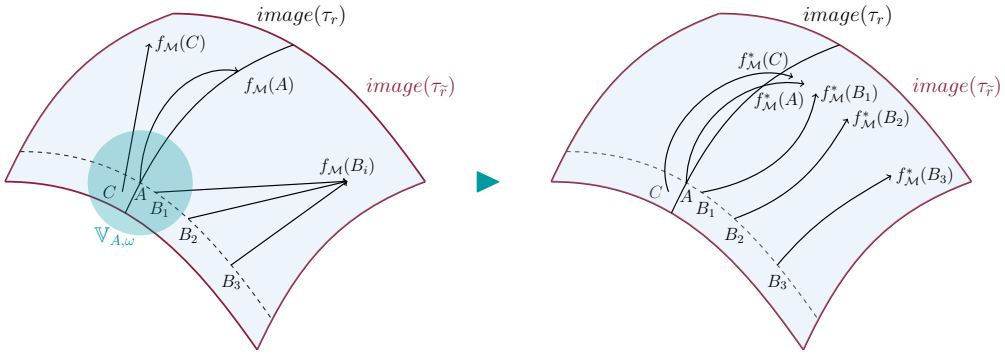

Figure 1: The schematic display of the unstable function $f_{\mathcal{M}}$ (left) and the averaged, stable $f^*_{\mathcal{M}}$ (right). In both pictures, the image of τ_r is depicted as black curve contained in the image of $\tau_{\widetilde{r}}$ shown as blue area (with magenta boundary). A is a rank r tensor, 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 tensor A) are mapped to the same point $f_{\mathcal{M}}(B_i)$. 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 a tensor (such as B_1 and C) is close to A, then this also holds for their function values. However, the $f^*_{\mathcal{M}}(A)$ is not rank r anymore (in fact, the image of $f^*_{\mathcal{M}}$ is generally not even rank $\widetilde{r}$).

we limit the disturbance that the left singular vectors U receive due to the variation of A to only one component (as limit case of $\sigma_1 \gg \sigma_2 \approx \omega$). From this, we will observe important consequences.

Lemma 2.2 (Averaged low rank matrix approximation)**.** *Let $\mathcal{M}$ be defined by* (1.1) *and $P = \mathcal{I}$. 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*

$$\widehat{f}_{\mathcal{M}}(A) := \frac{1}{|V_\omega|}\int_{V_\omega} f_{\mathcal{M}}\left((u_1 \mid u_2+\Delta u_2)\Sigma V^T\right)\mathrm{d}\Delta u_2$$

$$= \underbrace{u_1u_1^TM}_{\textit{optimization}} + \underbrace{(1-\alpha_\omega)^2u_2u_2^TM}_{\textit{regularization}} + \underbrace{\frac{2\alpha_\omega-\alpha_\omega^2}{m-2}(I_m-u_1u_1^T-u_2u_2^T)M}_{\textit{replenishment}} \tag{2.3}$$

$$V_\omega := \left\{\Delta u_2 \mid \|(u_1 \mid u_2+\Delta u_2)\Sigma V^T - A\| = \omega,\ (u_1 \mid u_2+\Delta u_2)\ \textit{is column orthogonal}\right\}$$

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 of the second singular vector, we obtain*

$$\frac{1}{|V_\omega|}\int_{V_\omega} f_{\mathcal{M}}\left((u_1 \mid \Delta u_2)\Sigma V\right)\mathrm{d}\Delta u_2 = u_1u_1^TM + \frac{1}{m-1}(I_m-u_1u_1^T)M,$$
$$\textit{where here}\quad V_\omega := \{\Delta u_2 \mid (u_1 \mid \Delta u_2)\ \textit{is column orthogonal}\}.$$

Proof. We parametrize V_ω. First, $\omega = \|(u_1 \mid u_2+\Delta u_2)\Sigma V^T - A\| = \|\Delta u_2\|\ \sigma_2$ and hence $\|\Delta u_2\| = \frac{\omega}{\sigma_2}$. 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

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

We integrate this over V_ω and obtain

$$\int_{V_\omega} f_{\mathcal{M}} = \int_{V_\omega} u_1u_1^TM + \int_{V_\omega}(1-\alpha_\omega)^2u_2u_2^TM + \int_{V_\omega}\Delta u_2^\perp{\Delta u_2^\perp}^TM$$

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 4.5 to

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

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 $HH^T = I_m - u_1u_1^T - u_2u_2^T$, since the rank of H is $m-2$ and $\mathit{range}(H) \perp \mathit{range}(U)$. 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 low 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. Note however that we fixed $\|\Delta u_2\| = \omega$ (for simplicity) as well as that for $\omega > \sqrt{2}\sigma$, Example 2.2 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**.
More importantly, the result $\widehat{f}_{\mathcal{M}}(A)$ in (2.3) is not low rank, yet is close to the rank 2 approximation $U(u_1^TM, (1-\alpha_\omega)^2u_2^TM))$, in which the first component U has remained the same. While the averaged model as in (2.2) remains the root idea, it appears too complicated to use for the derivation of a stable method $\mathcal{M}^*$. We instead consider a slightly modified approach in Section 4:

Lemma 2.3 (Low rank matrix approximation using a variational residual function)**.** *In the first situation of Example 2.2, we have*

$$\operatorname*{argmin}_{\widetilde{V}} \frac{1}{|V_\omega|}\int_{V_\omega} \|(u_1 \mid u_2+\Delta u_2)\widetilde{V} - M\|^2 \,\mathrm{d}\Delta u_2 = (u_1^T M \mid (1-\alpha_\omega)u_2^T M) \tag{2.4}$$

Proof. With the same derivation as in Lemma 2.2, 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 (2.3), we observe that there is only one difference, i.e. $(1-\alpha_\omega)^2$ has been replaced by $1-\alpha_\omega$. For our purpose, these terms are sufficiently similar if $\alpha_\omega \in (0,1)$.
The replenishment term however, which we so far ignored, is crucial. Without this term, some parts of the iterates will simply converge to zero for fixed ω (cf. Appendix B). We later bypass this problem by setting a lower limit to all occurring singular values.
We also refer to a Matlab implementation of a (superficially random) Monte Carlo approach to the unsimplified averaged micro-step $f^*_{\mathcal{M}}$ as in (2.2) for matrix completion. Likewise, an implementation of the final algorithm *SALSA* (Algorithms 2, 3), which is developed from the idea in Lemma 2.3, can be found for the matrix case, as well as for the tensor case of course, under `www.igpm.rwth-aachen.de/personen/kraemer`.

3 Notations and Reduction to Three Dimensions

As mentioned earlier, we reduce the d dimensional setting to a three dimensional one:

Notation 3.1 (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 and right unfolding $\mathfrak{L}(H) \in \mathbb{R}^{k_1\cdot n\times k_2}$, $\mathfrak{R}(H)\in\mathbb{R}^{k_1\times k_2\cdot n}$ by*

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

for $1\le j\le n$, $1\le \ell\le k_1$ and $1\le q\le k_2$. For a representation G, we correspondingly define the interface matrices

$$\begin{aligned}G^{<\mu} &= \mathfrak{L}(G_1\boxtimes\dots\boxtimes G_{\mu-1}) \in \mathbb{R}^{n_1\dots n_{\mu-1}\times r_{\mu-1}},\\ G^{>\mu} &= \mathfrak{R}(G_{\mu+1}\boxtimes\dots\boxtimes G_d) \in \mathbb{R}^{r_\mu\times n_{\mu+1}\dots n_d} \quad (\textit{cf. Definition 1.3}).\end{aligned}$$

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

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

For any representation it hence holds

$$(\tau_r(G))_{(\mu)} = G^{<\mu}\boxtimes G_\mu \boxtimes G^{>\mu}. \tag{3.1}$$

From now on we will mostly skip the symbol $\boxtimes$ in terms as in (3.1) or for any scalar product of a core and a matrix (where the matrix is regarded as scalar). This relation is displayed in Figure 2.

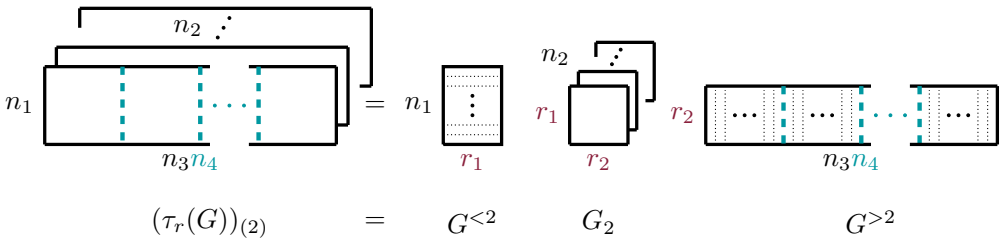

Figure 2: The decomposition of a core unfolding with respect to 2 of a four dimensional tensor into the left and right interface matrices as well as the intermediate core.

We will further use the following, convenient notations, since we often have to reshape, restrict or project objects.

Notation 3.2 (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.*

For example, for $P = \{p^{(i)} \mid i = 1, \dots, |P|\}$ and the rearrangement $(\cdot)_{(\mu)}$, which is used to summarize components $s = 1, \dots, \mu - 1$ as well as $s = \mu + 1, \dots, d$, let

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

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

$$P_{(\mu)}(j) = \{((p_1^{(i)}, \dots, p_{\mu-1}^{(i)}), (p_{\mu+1}^{(i)}, \dots, p_d^{(i)})) \mid p_\mu^{(i)} = j,\ i = 1, \dots, |P|\}. \tag{3.2}$$

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

W.l.o.g. 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. Remark 3.1):

Notation 3.3 (Reduction to three dimensions)**.** *When $\mu \in D$ 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* $S = P_{(\mu)}$

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

$$\begin{aligned} \mathcal{M}_r^{(\mu)}(G) &:= (G_1, \dots, G_{\mu-1}, G_\mu^+, G_{\mu+1}, \dots, G_d) \\ G_\mu^+ &:= \operatorname*{argmin}_{G_\mu} \|\tau_r(G) - M\|_P = \operatorname*{argmin}_{\widetilde{N}} \|L \cdot \widetilde{N} \cdot R - B\|_S \end{aligned} \tag{3.3}$$

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. In that sense, we only need to consider three-dimensional tensors

$A = \mathfrak{L}^{-1}(L) \boxtimes N \boxtimes \mathfrak{R}^{-1}(R)$ with mode size (n_L, n_N, n_R), rank (r_γ, r_θ) and singular values (γ, θ). For simplicity, we redefine τ_r for this case via $A = \tau_r(L, N, R)$.

4 Stable Alternating Least Squares Micro-Steps

Our motivation is to adapt the target function of each micro-step in order to obtain a stable method $\mathcal{M}^*$. One may construe a tensor as one test function. A micro-step of ALS then yields a minimizer only for this specific point. It is hence a reasonable approach to instead consider a set $V_\omega(A)$ of variations ΔA along the manifold of three dimensional TT-rank r tensors:

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

Let $A = \tau_r(L, N, R)$ and $(\Delta L, \Delta N, \Delta R)$ such that $A + \Delta A = \tau_r(L + \omega\Delta L, N + \omega\Delta N, R + \omega\Delta R)$. Then

$$\begin{aligned}\|\Delta A\|_F^2 &= \|(L + \omega\Delta L)(N + \omega\Delta N)(R + \omega\Delta R) - LNR\|_F^2 \\ &= \|\omega(\Delta LNR + L\Delta NR + LN\Delta R)\|_F^2 + \left(\mathcal{O}(\omega^2)\right)^2\end{aligned}$$

The term $\|\Delta LNR + L\Delta NR + LN\Delta R\|^2$ can be approximated, assuming the angles between the three summands are small[1], by $\|\Delta LNR\|^2 + \|L\Delta NR\|^2 + \|LN\Delta R\|^2$.

Definition 4.1 (Variational residual function)**.** *Let $\omega \geq 0,\ s_1, s_2 > 0$ and B, S, L, N, R as in Notation 3.3. We define the averaged residual function $C := C_{B,S,L,N,R}$ for $\mathbb{V}_\omega := \mathbb{V}_\omega(L, N, R)$ by:*

$$C(\widetilde{N}) := \int_{\mathbb{V}_\omega(L,N,R)} \|(L + s_1\Delta L)(\widetilde{N} + \Delta N)(R + s_2\Delta R) - B\|_S^2 \,\mathrm{d}\Delta L\,\mathrm{d}\Delta N\,\mathrm{d}\Delta R, \quad (4.1)$$

with

$$\mathbb{V}_\omega = \{(\Delta L, \Delta N, \Delta R) \mid \|\Delta LNR\|^2 + \|L\Delta NR\|^2 + \|LN\Delta R\|^2 \leq \omega^2\}$$

where s_1, s_2 are scalings that only depend on the proportions of the representation, to be specified by Lemma 5.1.

It is easy to see that 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 $\widetilde{N}$.

4.1 Standard Representation of a TT-Tensor

A representation $G = (L, N, R)$ can be changed without changing the generated tensor $A = \tau_r(G)$ ([20, 32]), more specifically

$$\tau_r(G) = \tau_r(\widetilde{G}) \quad \Leftrightarrow \quad \widetilde{G} = (\widetilde{L}, \widetilde{N}, \widetilde{R}) = (LT_1^{-1}, T_1NT_2^{-1}, T_2R) \quad (4.2)$$

for two regular matrices $T_1 \in \mathbb{R}^{r_\gamma \times r_\gamma}, T_2 \in \mathbb{R}^{r_\theta \times r_\theta}$. Using the unique TT-singular values, one can define a standard representation that is essentially unique (in terms of uniqueness of the truncated matrix SVD[2]). For the construction, an only slightly modified TT-SVD [28] is used.

[1] This is generically the case for large vectors v, w with uniformly distributed entries in $(-1, 1)$.

[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 $\{-1, 1\}$.

Lemma 4.2 (Standard representation)**.** *Let $A \in \mathbb{R}^{n_L \times n_N \times n_R}$ be a tensor. There exists an essentially unique representation (with minimal ranks)*

$$\mathcal{G} = (\mathcal{L}, \Gamma, \mathcal{N}, \Theta, \mathcal{R}) \tag{4.3}$$

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

Proof. uniqueness:
Let there be two such representations $\widetilde{\mathcal{G}}$ and $\mathcal{G}$. Since the left-singular vectors of $A^{(\{1\})}$ are essentially unique, we conclude $\widetilde{\mathcal{L}} = \mathcal{L}w_1$ for an orthogonal matrix w_1 that commutes with Γ. Via an SVD of $A^{(\{1,2\})}$ it follows that $\widetilde{\mathcal{R}} = w_2^T\mathcal{R}$ for an orthogonal matrix w_2 that commutes with Θ. Furthermore $\mathfrak{L}(\mathcal{L}\Gamma w_1\widetilde{\mathcal{N}}) = \mathfrak{L}(\widetilde{\mathcal{L}}\Gamma\widetilde{\mathcal{N}}) = \mathfrak{L}(\mathcal{L}\Gamma\mathcal{N})w_2$. The map $x \mapsto \mathfrak{L}(\mathcal{L}\Gamma x)$ is linear and, in this case, of full rank. This implies $\widetilde{\mathcal{N}} = w_1^T\mathcal{N}w_2$.
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 (4.2)). An SVD of $\widetilde{L}$ yields $\widetilde{L} = \mathcal{L}\ \Gamma\ V_1^T$, since $\mathcal{L}\ \Gamma\ \mathfrak{R}(V_1^T\ \widetilde{N}\ \widetilde{R})$ 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 $\mathfrak{L}(\mathcal{L}\ \widehat{N})\ \Theta\ (V_2^T\ \widetilde{R})$ is a truncated SVD of $A^{(\{1,2\})}$. We can finish the proof defining $\mathcal{N} := \Gamma^{-1}\ \widehat{N}$ and $\mathcal{R} = V_2^T\ \widetilde{R}$. Note that, by construction, $\mathfrak{L}(\Gamma\ \mathcal{N})$ is column-orthogonal.
implied orthogonality:
Using the essential uniqueness, it follows that $\mathfrak{L}(\Gamma\ \mathcal{N})$ must indeed be column-orthogonal. By analogously constructing the extended representation from right to left we would obtain that $\mathfrak{R}(\mathcal{N}\ \Theta)$ is row-orthogonal. By uniqueness it follows again that this is always the case. □

Remark 4.3 (Conventional form of standard representation)**.** *Throughout the rest of the article, the standard representation will mostly appear in form of a specific, conventional representation*

$$(L,\ N,\ R) = (\mathcal{L},\ \Gamma\ \mathcal{N}\ \Theta,\ \mathcal{R}), \tag{4.4}$$

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

4.2 Minimizer of the Averaged Residual Function

We define (from now on) our method as

$$\mathcal{M}^*(L,\ N,\ R) = (L,\ \operatorname*{argmin}_{\widetilde{N}} C(\widetilde{N}),\ R) \tag{4.5}$$

with $C = C_{B,S,L,N,R}$ as in (4.1). Although Theorem 4.7, or more specifically the regularity of $Y(j)$ given by (4.7), 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 $\mu \in \{1, d\}$ are derived from the general case (Remark 5.2).

Lemma 4.4 (Representation independent)**.** *The method $\mathcal{M}^*$ 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 (4.2), there exist two matrices T_1, T_2 such that

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

Hence

$$\widehat{C}(\widetilde{N}) = \int_{\widehat{\mathbb{V}}_\omega} \left\|(L + s_1\Delta\widehat{L}T_1^{-1})T_1\widetilde{N}T_2^{-1}(R + s_2T_2\Delta\widehat{R}) - B\right\|_S^2 \mathrm{d}\Delta\widehat{L}\,\mathrm{d}\Delta\widehat{N}\,\mathrm{d}\Delta\widehat{R},$$
$$\text{with } \widehat{\mathbb{V}}_\omega = \left\{(\Delta\widehat{L},\Delta\widehat{N},\Delta\widehat{R}) \mid \|\Delta\widehat{L}T_1^{-1}NR\|^2 + \|LT_1^{-1}\Delta\widehat{N}T_2R\|^2 + \|LNT_2\Delta\widehat{R}\|^2 \le \omega^2\right\}$$

The substitution $(\Delta\widehat{L},\Delta\widehat{N},\Delta\widehat{R}) \overset{\iota}{\to} (\Delta LT_1, T_1^{-1}\Delta NT_2, T_2^{-1}\Delta R)$ introduces a constant Jacobi Determinant $|\det(J_\iota)|$ for some J_ι. We obtain

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

The determinant is irrelevant to the minimizer and 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). □

Lemma 4.5 (Integral over all variations)**.** *Let $n,m \in \mathbb{N}$, $\omega \ge 0$ 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

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

Proof. The proof mainly works with symmetry arguments. See Appendix A for details. □

Corollary 4.6 (Integral over Kronecker product)**.** *Let $\omega_1 > 0$.*
Further, let $H \in \mathbb{R}^{(n_Xn_Y)\times(n_Xn_Y)}$ 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

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

for $(H^*)_{i,j} = tr(h_{i,j})$, $H = \sum_{i,j} h_{i,j}\otimes e_ie_j^T$, $h_{i,j} \in \mathbb{R}^{n_X\times n_X}$. For an analog $V_{\omega_2}^{(n_Y,m_Y)}$, $\omega_2 > 0$, we further have

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

Proof. Using the splitting $H = \sum_{i,j} h_{i,j}\otimes e_ie_j^T$, $h_{i,j} \in \mathbb{R}^{n_X\times n_X}$, Lemma 4.5 can be applied to each summand, separately for X and Y. □

We now derive the minimizer of the variational residual function (4.1). Due to Lemma 4.4, we can use the standard representation in form of Remark 4.3 for simplification. In this case, $\mathbb{V}_\omega$ takes the convenient form

$$\mathbb{V}_\omega(\mathcal{L},\ \Gamma\ \mathcal{N}\ \Theta,\ \mathcal{R}) = \{(\Delta L, \Delta N, \Delta R) \mid \|\Delta L \Gamma\|^2 + \|\Gamma \Delta N \Theta\|^2 + \|\Theta \Delta R\|^2 \leq \omega^2\}. \quad (4.6)$$

Theorem 4.7 (Minimizer of the ALS averaged residual function)**.** *Let* $(\mathcal{L},\ \Gamma,\ \mathcal{N},\ \Theta,\ \mathcal{R})$ *be the standard representation* (4.3) *for a tensor* A*. The minimizer* N^+ *of the residual function* $C_{B,S,\mathcal{L},\Gamma\mathcal{N}\Theta,R}$ *as in* (4.1) *is given by*

$$N^+(j) = \underset{\widetilde{N}(j)}{\operatorname{argmin}} \underbrace{\|\mathcal{L}\ \widetilde{N}(j)\ \mathcal{R} - B(j)\|^2_{S(j)}}_{\textit{standard ALS}} + \underbrace{\|Y(j)\ \textit{vec}(\widetilde{N}(j))\|^2_F}_{\textit{regularization}}, \quad j = 1, \ldots, n_N$$

where

$$Y(j) := \begin{pmatrix} \sqrt{n_L^{-1}\zeta_1}\ {\mathcal{R}_{-,S(j)_2}}^T \otimes \Gamma^{-1} \\ \sqrt{n_R^{-1}\zeta_2}\ \Theta^{-1} \otimes \mathcal{L}_{S(j)_1,-} \\ \sqrt{\rho_{1,2}\zeta_{1,2}}\ \Theta^{-1} \otimes \Gamma^{-1} \end{pmatrix} \quad (4.7)$$

with $S(j)_i = (x_i^{(1)}, x_i^{(2)}, \ldots)$*, for* $S(j) = x^{(1)}, x^{(2)}, \ldots$*,* $i = 1, 2$*. The constants* ζ, ρ *only depend on the proportions of the representation and sampling set (cf. Remark 4.9) as well as the constant scalings* s_1*,* s_2*.*

Proof. The proof is rather technical and can be found in Appendix A. □

This result may appear to be intricate. However, to calculate the minimizer is of the same order (with near same constant) as for standard ALS, for which the matrices $\mathcal{L}_{S(j)_1,-} \in \mathbb{R}^{a_j \times r_\gamma}$ and $\mathcal{R}_{-,S(j)_2} \in \mathbb{R}^{r_\theta \times a_j}$ $(a_j = |\{p \mid p \in P,\ p_\mu = j\}|)$ are required anyway (for further explanation, see (5.2),(5.3)). As an example, for the approximation of a fully available tensor, Theorem 4.7 reduces to the following.

Corollary 4.8 (Filter properties)**.** *For* $P = \mathcal{I}$*, the update is given by the so called filter*

$$\begin{aligned} \mathcal{F} &:= (I \otimes I + \zeta_1 \cdot I \otimes \Gamma^{-2} + \zeta_2 \cdot \Theta^{-2} \otimes I + \zeta_{1,2} \cdot \Theta^{-2} \otimes \Gamma^{-2})^{-1}, \quad (4.8) \\ N^+ &= \mathcal{F} \odot (\mathcal{L}^T\ B\ \mathcal{R}^T), \end{aligned}$$

where $\odot$ *acts matrix wise as Hadamard product.*

Proof. From $P = \mathcal{I}$, it follows that $\mathcal{R}_{-,S(j)_2}$ is an n_L-order copy of $\mathcal{R}$ and $\mathcal{L}_{S(j)_1,-}$ is an n_R-order copy of $\mathcal{L}$ (cf. (4.7)). Hence $Y(j) =: Y$ is independent of j. The minimizer $N^+(j)$ is given by

$$(K^TK)^{-1}K^T \begin{pmatrix} \operatorname{vec}(B(j)) \\ 0 \\ \vdots \end{pmatrix} \quad \text{for} \quad K = \left.\begin{pmatrix} \mathcal{R}^T \otimes \mathcal{L} \\ \sqrt{n_L^{-1}\zeta_1}\ \mathcal{R}^T \otimes \Gamma^{-1} \\ \vdots \\ \sqrt{n_R^{-1}\zeta_2}\ \Theta^{-1} \otimes \mathcal{L} \\ \vdots \\ \sqrt{\zeta_{1,2}}\ \Theta^{-1} \otimes \Gamma^{-1} \end{pmatrix}\right\} \begin{matrix} n_L\text{-times} \\ \\ \\ n_R\text{-times} \\ \\ \\ \end{matrix}.$$

The factors $\sqrt{n_L^{-1}}$ and $\sqrt{n_R^{-1}}$ vanish in K^TK due to the multiple rows involving the orthogonal matrices $\mathcal{R}$ and $\mathcal{L}$. Furthermore, $(K^TK)^{-1}$ is diagonal, such that equation can be restated using the Hadamard product. □

Remark 4.9 (Specification of constants (cf. Appendix A))**.** *Let $\#_{\mathcal{R}} := size(\mathcal{R})$, $\#_{\mathcal{N}} := size(\mathcal{N})$, $\#_{\mathcal{L}} := size(\mathcal{L})$ be the sizes of the tensor components. The constants in Theorem 4.7 are given by $\rho_{1,2} = |S(j)|n_L^{-1}n_R^{-1} = |S(j)|(|\mathcal{I}|/n_N)^{-1}$ and*

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

We finish this subsection with the central theoretical statement of this paper. The Tensor Restricted Isometry Property (e.g. [30]) does not hold for any non trivial sampling set $P \subsetneq \mathcal{I}$. We however only need to work with a modified version as follows.

Definition 4.10 (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, if there exist $0 \le c < 1$ and $\rho > 0$ with*

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

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

Note that the constants are independent of the specific, chosen representation.

Lemma 4.11 (Likelihood of the iTRIP)**.** *Let $\mathcal{T}$ be the subset of 3 dimensional tensors with rank $r = (r_\gamma, r_\theta)$. Let P be a (random) sampling that fulfills $|S(j)| \ge r_\gamma r_\theta$ for all $j = 1, \dots, n_N$. Then almost every $A \in \mathcal{T}$ has the iTRIP. If for one j, $|S(j)| < r_\gamma r_\theta$, then no $A \in \mathcal{T}$ 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, $(R^T \otimes L)_{\mathrm{vec}(S(j)),-}$ has full rank for each j. Due to the provided slice density of P, each matrix $(\mathcal{R}^T \otimes \mathcal{L})_{\mathrm{vec}(S(j)),-}$ 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 (in the matrix case for example, some columns or rows may indeed have very few samples). If the iterate is close to such a tensor, the likelihood grows to encounter overfitting (cf. Example 1.2), but the regularization (4.7) already compensates this.

Lemma 4.12 (Partial matrix inverse by divergent parts)**.** *For a partition[3] of indices $\{1, \dots, n\}^2 = (\omega_1 \cup \complement\omega_1) \times (\omega_2 \cup \complement\omega_2)$, we define $\Omega := \omega_1 \times \omega_2$ and $\Omega^c := \complement\omega_1 \times \complement\omega_2$. Let $\{A^{(k)}\}_k$, $\{J^{(k)}\}_k \subset \mathbb{R}^{n\times n}$ be series of symmetric matrices, $supp(J^{(k)}) \subset \Omega$.*
If $lim_{k\to\infty} A^{(k)}|_{\Omega^c} = A|_{\Omega^c}$, $A|_{\Omega^c}$ s.p.d, and $\sigma_{min}(J^{(k)}|_\Omega) \to \infty$, then $B := lim_{k\to\infty}(A^{(k)} + J^{(k)})^{-1}$ exists and we have $B|_{\Omega^c} = (A|_{\Omega^c})^{-1}$ and $B|_{\complement\Omega^c} = 0$.

[3]The symbol $\complement$ denotes the set-complement to a (by context) given set W, i.e. $\complement s = W \setminus s$.

Proof. First, w.l.o.g., let $\Omega = \{m+1, \dots, n\}^2$. Otherwise we can apply permutations. Further, let $B^{(k)} := A^{(k)} + J^{(k)}$. 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 $A_{1,1}^{(k)} = B_{1,1}^{(k)}$ and $A_{1,1} = A|_{\Omega^c}$ is s.p.d, $A_{1,1}^{(k)}$ is invertible for all $k > K$ for some K and hence $\lim_{k\to\infty}(B_{1,1}^{(k)})^{-1} = A_{1,1}^{-1}$. Further, $\sigma_{\min}(B_{2,2}^{(k)}) > \sigma_{\min}(J_{2,2}^{(k)}) - \sigma_{\max}(A_{2,2}^{(k)}) \to \infty$ and hence $\|(B_{2,2}^{(k)})^{-1}\| \to 0$. Therefore, for $k > \widetilde{K}$ and $H^{(k)} := B_{1,1}^{(k)} - B_{1,2}^{(k)}(B_{2,2}^{(k)})^{-1}(B_{1,2}^{(k)})^T$, it is $\sigma_{\min}(H^{(k)}) > \sigma_{\min}(A_{1,1})/2$. By block-wise inversion of $B^{(k)}$, it then follows $((B^{(k)})^{-1})_{1,1} = (H^{(k)})^{-1} \to (A_{1,1}^{(k)})^{-1}$. Similarly, $((B^{(k)})^{-1})|_\Omega \to 0$. □

One last step remains, since we cannot allow ζ to depend on the rank r. For now, we redefine the method $\mathcal{M}^*$ to directly yield the result in Theorem 4.7 for arbitrary constants ζ, i.e.

$$\mathcal{M}_\zeta^*(L,\ N,\ R) := (L,\ N^+,\ R). \tag{4.9}$$

We explain in Section 5 and Lemma 5.1 how the scalings s_1, s_2 as well as ω are used to obtain one specific $\mathcal{M}_\zeta^*$ from $\mathcal{M}^*$, for which ζ is indeed independent of r. It is easy to see that $\mathcal{M}_\zeta^*$ is (trivially) representation independent as it is defined via the essentially unique standard representation.

Theorem 4.13 (Stability of the method $\mathcal{M}_\zeta^*$)**.** *Let B be the target tensor, S the sampling set, arbitrary but fixed and $\mathcal{M}_\zeta^*$ as in* (4.9)*.*

- ($\omega = 0$) *The unregularized method* (3.3) *is stable for fixed rank at all points A^* that have the iTRIP (cf. Def. 4.10).*
- ($\omega > 0$) *The regularized method $\mathcal{M}_\zeta^*$ as defined by* (4.9) *(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 fix-rank stable).*

Proof. Let A^* be a fixed tensor with TT-ranks r^*.
1. fix-rank stability
We first show that $\mathcal{M}^*$ is stable for fixed rank. Let A_i be a sequence with $\operatorname{rank}(A_i) = r^*$ and $A_i \to A^*$. Let $\mathcal{G}^* = (\mathcal{L}^*, \Gamma^*, \mathcal{N}^*, \Theta^*, \mathcal{R}^*)$ be the standard representation of A^* as well as $\mathcal{G}_i$ 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 $\ell_1, \dots, \ell_L$. Since $A_i \to A^*$, their singular values also converge (e.g. [38]). We can hence conclude from [9,36] that there exist sequences of block diagonal, orthogonal matrices W_i and M_i with block sizes $k_1, k_2 - k_1, \dots, k_K - k_{K-1}$ and $\ell_1, \ell_2 - \ell_1, \dots, \ell_L - \ell_{L-1}$, respectively, such that

$$\|\mathcal{L}_i W_i - \mathcal{L}^*\|_F \to 0 \quad \text{and} \quad \|M_i \mathcal{R}_i - \mathcal{R}^*\|_F \to 0, \tag{4.10}$$

since the standard representation includes left and right singular vectors. We have to show that the tensors $Z_i = \tau_r(\mathcal{L}_i, N_i, \mathcal{R}_i) = \tau_r(\mathcal{M}^*(\mathcal{L}_i, \Gamma_i \mathcal{N}_i \Theta_i, \mathcal{R}_i))$ converge to the analogously defined Z^*. For fixed j, we define for each single $\mathcal{G}_i$ the matrix $Y_i = Y(j)$ from Theorem 4.7 and $z_i := (\mathcal{R}_i^T \otimes \mathcal{L}_i)$ such that

$$N_i(j) = \operatorname*{argmin}_{\widetilde{N}(j)} \left\| \begin{pmatrix} (z_i)_{\mathrm{vec}(S(j)),-} \\ Y_i \end{pmatrix} \mathrm{vec}(\widetilde{N}(j)) - \begin{pmatrix} \mathrm{vec}(B(j))|_{\mathrm{vec}(S(j))} \\ 0 \end{pmatrix} \right\|, \tag{4.11}$$

$\mathrm{vec}(Z_i(j)) = z_i\ \mathrm{vec}(N_i(j))$.

We define the shifted matrices

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

Due to (4.10), 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 (4.11), we obtain

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

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) $w = 0$: 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

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

This directly yields convergence of $(Z_i(j)) \to (Z^*(j))$ since all involved factors converge.
(ii) $w > 0$: Here, we use that $\sigma_{\min}(Y^*) > 0$ and $\sigma_{\min}(z^*_{\mathrm{vec}(S(j)),-}) \geq 0$. We then obtain convergence since

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

This proves fix-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 4.12, let Ω be the smallest cross product set, such that $(Y_i^{M,W^T}\ Y_i^{M,W})|_{\Omega^c}$ 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

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

and

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

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

5 Results Transferred Back to a d-Dimensional Tensor

In this Section, we return to a d-dimensional tensor. In Remark 4.9, we have $\#_{\mathcal{R}} = size(\mathcal{R}) = r_\theta \prod_{i=s+1}^{d} n_i$, $\#_{\mathcal{N}} = size(\mathcal{N}) = r_\gamma n_s r_\theta$, $\#_{\mathcal{L}} = size(\mathcal{L}) = r_\gamma \prod_{i=1}^{s-1} n_i$. By combining modes (cf. Notation 3.3), the *sizes* of the left as well as right side have been drastically overrated and distorted, considering that the degrees of freedom of $\mathcal{L} = \mathcal{G}^{<s}$ and $\mathcal{R} = \mathcal{G}^{>s}$ are given by a sum, not a product, of the degrees of freedom of the single modes. We choose one of the few remaining options through which the method becomes stable. We artificially set

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

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 micro-step. Furthermore, the near common parts of the denominators, $\#_{\mathcal{R}} + \#_{\mathcal{N}} + \#_{\mathcal{R}} + 2(+2)$, can be incorporated into ω^2, so we omit them in the following sense:

Lemma 5.1 (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,

$$\zeta_1^{(\mu)} = \omega^2 \frac{\sum_{s=1}^{\mu-1} n_s}{\sum_{s=1}^{d} n_s}, \quad \zeta_2^{(\mu)} = \omega^2 \frac{\sum_{s=\mu+1}^{d} n_s}{\sum_{s=1}^{d} n_s}, \quad \zeta_{1,2}^{(\mu)} = \zeta_1^{(\mu)} \zeta_2^{(\mu)} (1 + \mathcal{O}(E^{-1})). \tag{5.1}$$

Proof. See appendix A. □

The value E^{-1} is in general far below machine accuracy, such that we (from now on) ignore the factor $(1 + \mathcal{O}(E^{-1}))$. 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 4.13. In this context, for fixed μ, the matrices $\mathcal{L}_{S(j)_1,-} \in \mathbb{R}^{a_j \times r_\gamma}$ and $\mathcal{R}_{-,S(j)_2} \in \mathbb{R}^{r_\theta \times a_j}$, $a_j = |\{p \mid p \in P,\ p_\mu = j\}| = |P_{(\mu)}(j)|$ (cf. (3.2)), are given by

$$\left(\mathcal{L}_{S(j)_1,-}\right)_{\ell,-} = G_1(p_1^{(i_\ell)}) \cdot \ldots \cdot\ G_{\mu-1}(p_{\mu-1}^{(i_\ell)}), \tag{5.2}$$
$$\left(\mathcal{R}_{-,S(j)_2}\right)_{-,\ell} = G_{\mu+1}(p_{\mu+1}^{(i_\ell)}) \cdot \ldots \cdot\ G_d(p_d^{(i_\ell)}), \quad \ell = 1, \ldots, a_j,\ p^{(i_\ell)} \in P_{(\mu)}(j), \tag{5.3}$$

for a representation G for which $\mathcal{L} = G^{<s}$ and $\mathcal{R} = G^{>s}$.

Remark 5.2 (Case $\mu = 1, d$)**.** *For $\mu = 1, d$ in Theorem 4.7, the same formula can be used by formally setting $G^{<1} = \mathcal{L} = 1$, $G^{>d} = \mathcal{R} = 1$ and $\zeta_2^{(1)} = 0$, $\zeta_1^{(d)} = 0$, $\zeta_{1,2}^{(1),(d)} = 0$, respectively.*

Since all micro-steps $\mathcal{M}^*$ are stable, we call this regularized ALS method stable - hence the name SALSA (Stable ALS Approximation). We summarize in Algorithm 1 one full left sweep $\mu = 1 \to d$ of SALSA for some fixed rank r. Note that the algorithm remains with the same order of computational complexity $\mathcal{O}(dr^4|P|)$, and near same constants. The simpler matrix case ($d = 2$) is carried out in Algorithm 2.

Algorithm 1 SALSA Sweep

set $\sigma_0 \equiv \sigma_d \equiv 1$
Require: limits $\sigma_{\min}^{(\mu)}$, parameter ω, initial guess $A = \tau_r(G)$ for which $\mathfrak{R}(G_2), \dots, \mathfrak{R}(G_d)$ are row-orthogonal and data points $M|_P$
for $\mu = 1, \dots, d$ **do**
  compute the SVD $U\Sigma^{(\mu)}V^T := \mathfrak{L}(G_\mu)$ and update $\sigma^{(\mu)}$
  update $G_{\mu+1} := V^T G_{\mu+1}$ and G_μ via $\mathfrak{L}(G_\mu) = U\Sigma^{(\mu)}$
  set $\zeta_1^{(\mu)}, \zeta_2^{(\mu)}, \zeta_{1,2}^{(\mu)}$ as defined by (5.1)
  for $j = 1, \dots, n_\mu$ **do**
    compute the update $N_\mu(j)$ from the least squares problem given by Theorem 4.7 for $\mathcal{L} = G^{<s}$, $\mathcal{R} = G^{>s}$ (cf. Remark 5.2)
  end for
  if $\mu \neq 1$ **then**
    compute the SVD $U\widetilde{\Sigma}V^T := \mathfrak{R}(N)$
    update N via $\mathfrak{R}(N) := U\ \mathrm{diag}(\{\max(\widetilde{\sigma}_i, \sigma_{\min}^{(\mu-1)}) \mid i = 1, \dots, r_\mu\})\ V^T$
  end if
  if $\mu \neq d$ **then**
    compute the SVD $U\widetilde{\Sigma}V^T := \mathfrak{L}(N)$
    update $\sigma_i^{(\mu)} := \max(\widetilde{\sigma}_i, \sigma_{\min}^{(\mu)}),\ i = 1, \dots, r_\mu$
    update N via $\mathfrak{L}(N) := U$
    set $G_{\mu+1} := \Sigma^{(\mu)}V^T G_{\mu+1}$
  end if
  update $G_\mu := N$
end for

6 Semi Implicit and Non Uniform Rank Adaption

The stabiliy of SALSA is used to establish an in principle simple rank adaption. For a more detailed analysis and motivation, we refer to Appendix B. We capture the magnitude of regularization caused by the individual singular vectors $\sigma^{(\mu)}$:

Definition 6.1 (Minimal filter values)**.** *Define the entries of $F^{(\mu)} \in (0,1)^{r_\mu}$ via*

$$F_i^{(\mu)} := \max(\mathcal{F}_{1,i}^{(\mu)}, \mathcal{F}_{i,1}^{(\mu+1)}), \tag{6.1}$$

where $\mathcal{F}^{(0)} = \mathcal{F}^{(d)} := 0$ and $\mathcal{F}^{(\mu)}$ (for each μ) is defined in Corollary 4.8.

This magnitude is then used to define certain thresholds for all singular values.

Definition 6.2 (Virtual ranks and virtual singular values)**.** *Let $0 < F_{virt} < F_{stab} < 1$ be fixed. A singular value $\sigma_i^{(\mu)}$ is called virtual, if $F_i^{(\mu)} < F_{virt}$ and denoted stabilized*

Algorithm 2 Stable Matrix Completion

Require: limit $\sigma_{\min}$, parameter ω, initial guess $A = XY^T \in \mathbb{R}^{n_1 \times n_2}$ such that Y contains the right singular vectors of A and data points $M|_P$

for $i = 1, \dots, n_1$ update

$$X := \underset{\widetilde{X}_{i,-}}{\operatorname{argmin}} \|\widetilde{X}_{i,-} Y^T - M_{i,-}\|^2_{P_{i,-}} + \frac{|P_{i,-}|}{n_2} \frac{\omega^2 n_1}{n_1 + n_2} \|\widetilde{X}_{i,-} \Sigma^{-1}\|_F^2 \tag{5.4}$$

compute the SVD $U\Sigma V^T := X$ and update $\sigma_i := \max(\widetilde{\sigma}_i, \sigma_{\min}),\ i = 1, \dots, r$

set $X := U$ and $Y^T := \Sigma V^T Y^T$

for $i = 1, \dots, n_2$ update

$$Y_{-,i} := \underset{\widetilde{Y}_{-,i}}{\operatorname{argmin}} \|X \widetilde{Y}^T_{-,i} - M_{-,i}\|^2_{P_{-,i}} + \frac{|P_{-,i}|}{n_1} \frac{\omega^2 n_2}{n_1 + n_2} \|\Sigma^{-1} \widetilde{Y}^T_{-,i}\|_F^2 \tag{5.5}$$

compute the SVD $U\Sigma V^T := Y$ and update $\sigma_i := \max(\widetilde{\sigma}_i, \sigma_{\min}),\ i = 1, \dots, r$

set $X := XU\Sigma$ and $Y^T := V^T$

(with respect to F_{stab}) if $F_i^{(\mu)} > F_{stab}$. The virtual rank of $A = \tau_r(G)$ is given by its exact rank $r = r(A)$, while the stabilized rank only includes the stabilized singular values.

The trick is to overestimate all ranks by 1 and to gradually decrease ω (as well as the singular value limit). During several iterations, each last singular value $\sigma^{(\mu)}_{r_\mu}$ just equals $\sigma^{(\mu)}_{\min}$ (cf. Algorithm 3). It does thereby only marginally influence the optimization, which is why we use the term *virtual*. However, at a certain point, the according singular values exceed the minimum and then stabilize. Each time this happens and certain criteria hold, the technical rank is increased by 1 (by adding a virtual singular value using random terms). Vice versa, a rank is cut if the stabilized rank is by 2 lower than the virtual rank. The rest of this subsection will deal with remaining details.

Definition 6.3 (Control set)**.** *For a given index set P, we define $P_2 \subset P$ as control set. This set may be chosen randomly or specifically distributed as well. The actual set used for the optimization is replaced by $P \leftarrow P \setminus P_2$ (keeping the same symbol).*

It is not easy to give a general criterion when to terminate the algorithm. Often, an estimate for an upper limit to all ranks provides an efficient criterion. We here measure the improvement between rank increases, but there might be more suitable approaches.

Remark 6.4 (Blocking rank increases)**.** *For every previously taken value k, let $\widehat{G}^{(k)}$ be the representation given immediately before the k-th rank increase. Set*

$$R_X(k+1) := \|\tau_r(\widehat{G}^{(k)}) - M\|_X, \quad X \in \{P, P_2\}.$$

Define $\beta_X = \left|1 - \frac{\|\tau_r(G) - M\|_X}{R_X(\sum_{i=1}^{d-1} r_i - (d-1))}\right|$ for the current representation G.

As long as $\sum_{i=1}^{d-1} r_i \geq 2(d-1)$ and one of the following criteria is fulfilled, rank increases are blocked:

- $\beta_P < \beta_{\min}$
- $\beta_{P_2} < \beta_{\min}$ *(cf. Definition 6.3)*
- $\sum_{s=1}^{d} r_{s-1} r_s n_\mu - \sum_{i=1}^{d} r_\mu^2 > |P|/1.2$ *(degrees of freedom too high)*

Definition 6.5 (Unblocked ranks)**.** *We define*

$$\mathcal{U} = \begin{cases} \emptyset & \text{if rank increases are blocked (cf. Remark 6.4)} \\ \{s \in \{2, \dots, d\} \mid r_\mu + 1 \leq \min(n_s r_{s-1}, n_{s+1} r_{s+1}, r_{lim})\} & \text{otherwise} \end{cases}$$

where $r_{lim} \in \mathbb{N}$ *is a given, technical limit to any rank.*

Remark 6.6 (Decline of ω)**.** *Let* $G^{(\texttt{iter})}$ *be the representation after iteration number* $\texttt{iter} = 1, 2, \dots$*. Define*

$$\gamma_X^i := \frac{\mathrm{Res}_X(G^{(i)})}{\mathrm{Res}_X(G^{(i-1)})}, \qquad i = \texttt{iter} - 4 \dots \texttt{iter},\ X \in \{P, P_2\}$$

the arithmetic mean of the last 5 *residual reduction factors for the sampling and control residual. We say* ω *is minimal, if there exists a stabilized rank equal to* r_{lim} *or if* $\mathcal{U} = \{\}$ *(Definition 6.5) and all ranks are stabilized with respect to* $\widetilde{F}_{stab}$ *for a fixed* $F_{stab} < \widetilde{F}_{stab}$ *that is close to, yet less than* 1*. The parameter is regulated as follows: Initialize* $\widetilde{\omega} = \omega_0$*. After each iteration* `iter`*, if*

- ω *is not minimal and if either*
  - *the singular spectrum does not currently change too much and*
  - $\gamma_P^i < \gamma^*$ *or* $\gamma_{P_2}^i < \gamma^*$

  or
  - $\mathrm{Res}_P(G^{(\texttt{iter})}) > \mathrm{Res}_P(G^{(\texttt{iter}-1)})$

then $\widetilde{\omega}$ *is decreased by a constant factor of* f_ω*. Set then* $\omega = \widetilde{\omega} \|\tau_r(G^{(\texttt{iter})})\|_{\mathcal{I}}$*.*

Remark 6.7 (Changing ranks)**.** *The* μ*-th rank is increased if the following conditions hold:*

- $\mu \in \mathcal{U}$ *(Definition 6.5)*
- $\widetilde{\omega}$ *has been decreased in the previous iteration*
- $\sigma_{r_\mu}^{(\mu)}$ *is stabilized*

The representation is then expanded randomly, such that for the new singular value holds $\sigma_{r_\mu+1}^{(\mu)} = \sigma_{min}^{(\mu)}$ *and all other singular values remain equal.*
If, in contrast, at any time $\sigma_{r_{\mu-1}}^{(\mu)}$ *is virtual (and hence* $\sigma_{r_\mu}^{(\mu)}$ *as well), the rank is decreased by* 1 *and the tensor truncated.*

By this kind of rank adaption, only virtual singular values are ever introduced or removed. This is to be understood as the main idea behind SALSA. The exact rank is not relevant anymore within the optimization, only the magnitude of ω compared to the singular values matters.

Remark 6.8 (Termination)**.** *Let* $i^* = argmin_i \mathrm{Res}_{P_2}(G^{(i)})$ *and* $f_{P_2} > 1$ *be fixed. If one of the following criteria holds, then the algorithm terminates.*

- ω *is minimal (Remark 6.6) (convergence)*
- $\texttt{iter} > 10$ *and* $\mathrm{Res}_{P_2} > f_{P_2} \cdot \mathrm{Res}_{P_2}(G^{i^*})$ *(Definition 6.3) (divergence)*

As final result, G^{i^*} *is chosen (it may be cut to its stabilized rank).*

It remains to substitute the replenishment term in (2.3) in order to prevent virtual singular values from quickly converging to zero. Otherwise, they become essentially invisible to the algorithm and are not be picked up in subsequent steps.

Definition 6.9 (Singular value limit)**.** *The lower limit to the singular values is defined as fixpoint of*

$$\sigma_{min}^{(\mu)} \mapsto \frac{1}{\sum_{\mu=1}^d n_\mu} \left(1 - F_{min}^{(\mu)}(\sigma_{min}^{(\mu)})\right) \mathrm{Res}^{est} \tag{6.2}$$

where $F^{(\mu)}_{min}(\sigma^{(\mu)}_{min})$ is defined the same way as $F^{(\mu)}$ (see (6.1), (4.8)), but assuming that all last singular values equal the minimal $\sigma^{(\mu)}_{min}$. The value $\mathrm{Res}^{est} > 0$ is a pessimistic estimator for the full residual,

$$\mathrm{Res}^{est} := (\sqrt{|\mathcal{I}|/|P_2|}\mathrm{Res}_{P_2})^{3/2}\ (\sqrt{|\mathcal{I}|/|P|}\mathrm{Res}_P)^{-1/2}.$$

In practice, it is sufficient to perform a damped fixpoint iteration parallel to the decreases of $\widetilde{\omega}$ to obtain $\sigma^{(\mu)}_{min}$. Furthermore, the decrease of $\widetilde{\omega}$ is accelerated if $F^{(\mu)}_{virt}$ is much lower.

6.1 The SALSA Algorithm

We summarize the previous results in Algorithm 3. For the technical realizations, we refer to Section 7 and for the explicit choices of tuning parameters, see Subsection 7.3. The Matlab implementation (as well as all programs necessary to produce the results in Section 7) can further be found under `www.igpm.rwth-aachen.de/personen/kraemer`. The order of computational complexity does not exceed $\mathcal{O}(dr^4\#P)$, 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.

Algorithm 3 SALSA Algorithm

Require: $P \subset \mathcal{I}$, $M|_P$ (and parameters)
  initialize G s.t. $\tau_r(G) \equiv$ const, $|P|\|\tau_r(G)\|_F^2 = |\mathcal{I}|\|M|_P\|_P^2$ for $r \equiv 1$ and $\widetilde{\omega} = 1/2$
  split off a small control set $P_2 \subset P$ (Definition 6.3)
  proceed one or a few ordinary ALS sweeps (Algorithm 1 for $\omega \equiv 0$)
  for `iter` $= 1, 2, \ldots$ **do**
    ONCE: after a few iterations, introduce virtual ranks ($\Rightarrow r \equiv 2$)
    proceed SALSA sweep* (Algorithm 1)
    *: decrease $\widetilde{\omega}$ if progress low (Remark 6.6 applies)
    if *: a singular value becomes stabilized/virtual (Remark 6.7 applies) **then**
      increase/decrease the virtual rank
    end if
    if final breaking criteria apply (Remark 6.8) **then**
      terminate algorithm
    end if
  end for

7 Numerical Experiments

We consider the following two algorithms:

- standard ALS (Algorithm 1 for $\omega \equiv 0$)
- SALSA (Algorithm 3)

We explain how ranks are adapted for ALS in Section 7.1, give details for data acquisition and measurements in Section 7.2 as well as tuning parameters in Section 7.3. We analyze the results in the latter Section 7.8.

For each test, we give a (too large) upper bound r_{lim} for the maximal rank of the iterates. We like to emphasize that, in contrast to rank adaption itself, such a bound can subsequently be increased if this might yield improvements - since this does only pose a one dimensional problem. Such a limit is not obligatory, but in specific cases the

necessarily coarse criteria in Remark 6.4 only hold for very large rank, such that the algorithms would use up a lot of time without changing the results. For simplicity, we use a common mode size $n = n_1 = \ldots, n_d$.

7.1 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

$$\begin{aligned} T &:= (G^{<\mu-1})^T \ ((M - \tau_r(G))|_P)_{(\mu-1,\mu)} \ (G^{>\mu})^T, \\ \alpha_{i_{\mu-1},i_\mu} &= \operatorname*{argmin}_{\widetilde{\alpha}_{i_{\mu-1},i_\mu}} \| G^{<\mu-1} \ (G_{\mu-1}(i_{\mu-1}) \cdot G_\mu(i_\mu) \\ &\qquad + \widetilde{\alpha}_{i_{\mu-1},i_\mu} T(i_{\mu-1}, i_\mu)) \ G^{>\mu} - M_{(\mu-1,\mu)} \|_{P_{(\mu-1,\mu)}(i_{\mu-1},i_\mu)}. \end{aligned}$$

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 $\mathfrak{H} \in \mathbb{R}^{r_{\mu-2} n_{\mu-1} \times r_\mu n_\mu}$. Then $\sigma_+^{(\mu)} := \|\mathfrak{H}\|_2$, the largest singular value of $\mathfrak{H}$. This approach is very similar to two-fold DMRG micro-steps as defined in [19], but a bit more regularized. The corresponding rank $\mu = \operatorname{argmin}_{\widetilde{\mu}} \sigma_+^{(\widetilde{\mu})}$ is increased by 1, using a rank 1 approximation of $\mathfrak{H}$. Basically the same termination criteria as for SALSA are used, although some criteria that are based on ω are replaced as well as possible. 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.

7.2 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 $\mu = 1, \ldots, d$ and each index $i_\mu \in \mathcal{I}_\mu$ we pick $C_{sf} \cdot r_P^2$ indices $i_1, \ldots, i_{\mu-1}, i_{\mu+1}, \ldots, i_d$ at random (uniformly). This gives in total $|P| \lesssim C_{sf} \cdot dnr_P^2$ samples (excluding duplicate samples). The rank r_P is artificial, such that C_{sf} can be interpreted as sampling factor. After all, the degrees of freedom of a TT-tensor of common rank r is slightly less than dnr^2. As a verification set C, we use a set of the same cardinality as P that is generated in the same way. *Order of optimization:* Instead of the sweep we gave before ($\mu = 1, \ldots, d$) for simplicity, we alternate between two sweeps ($\mu = 1, \ldots, h, \quad \mu = d, \ldots, h, \quad h = \lfloor d/2 \rfloor$) to enhance symmetry. *Averaging:* With $\langle\cdot\rangle_{\mathrm{ar}}$ we denote the arithmetic mean and by $\langle\cdot\rangle_{\mathrm{geo}}$ the geometric mean which we use for logarithmic scales.

7.3 Implementation Details and Tuning Parameters

All tests were done using a (pure) Matlab implementation, so the time performances should be evaluated carefully. Section 6 involves several parameters and relations to enable a full understanding of the black box algorithm. These have been chosen equally for all experiments with respect to best results, not speed, and could be relaxed for easier problems (or in practice for first trials) to reduce timing considerably. It shall hence be mentioned in advance that the number of iterations for the regularized algorithms is in general much higher. Straightening the tolerances for ALS (hence allowing more

iterations) however, does not lead to notable improvements, or even the opposite.
The parameters are given by: $\gamma^* = 10^{-3}$, $f_\omega = 1.1$, $F_{\text{virt}} := 0.33$, $F_{\text{stab}} := 0.99$, $\widetilde{F}_{\text{stab}} := 0.999$, $\beta_{\min} := 0.02$, $f_{P_2} := 2.5$, $|P_2|/|P| = 1/20$. The specific choices are heuristic (based on experience), but likewise recommendable for other problems. We observed that any reasonable values near these 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.

7.4 Approximation of a Tensor with Near Uniform Singular Spectrum

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

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

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. The

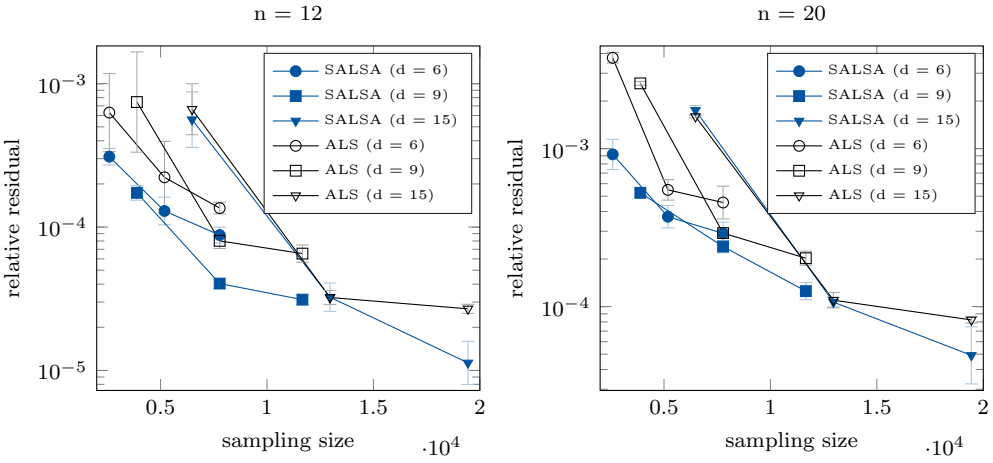

Figure 3: ($d = 6, 9, 15$, $r_P = 6$, $r_{lim} = 10$, $n = 12, 20$, $C_{sf} = 2, 4, 6$) Plotted are, for varying dimension and mode size, the averaged relative residuals $\langle R_C / \|M_C\| \rangle_{\text{geo}}$ and accordant standard deviations as functions of the sampling size $|P|$ as result of 20 trials, for ALS (black) and SALSA (blue, filled symbols).

results are plotted in Figure 3 (see Appendix C for Table 1).

7.5 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 generic tensors:

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

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

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

In contrast to the tensor in Section 7.4, 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 Figure 4 (see Appendix C for Table 2).

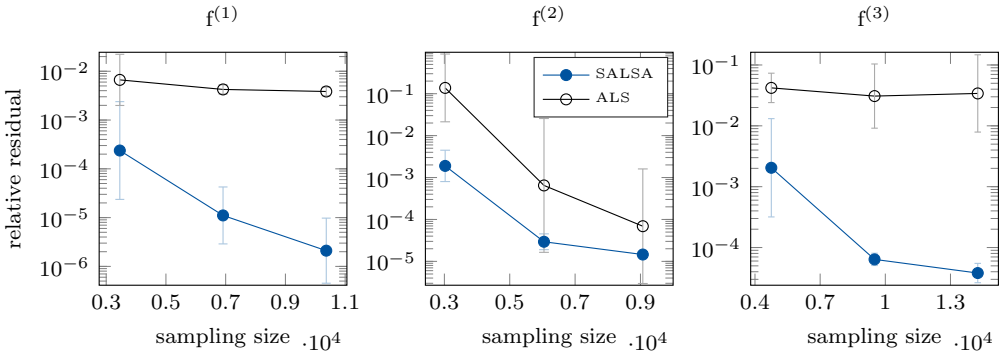

Figure 4: ($d_1 = 8, d_2 = 7, d_3 = 11$, $r_P = 6$, $r_{lim} = 10$, $n = 8$, $C_{sf} = 2, 4, 6$) Plotted are, for the tensors $f^{(1)}$ (left), $f^{(2)}$ (middle) and $f^{(3)}$ (right), the averaged relative residuals $\langle R_C/\|M_C\|\rangle_{\text{geo}}$ and accordant standard deviations as functions of the sampling size $|P|$ as result of 20 trials, for ALS (black) and SALSA (blue, filled symbols).

7.6 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[4]. Here it is required to set $\beta_{\min} = 0$. The ranks are generated randomly, but it is ensured that $\langle r\rangle_{\text{ar}} \geq 2/3k$ 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 $G_1,\dots,G_d$ a uniformly distributed random value in $[-0.5, 0.5]$. Subsequently, the singular values $\Sigma^{(1)},\dots,\Sigma^{(d-1)}$ are forced to take uniformly distributed random values in $[0,1]$ (up to scaling). This is achieved by successive replacements of the

[4]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 is **no indication** yet that the sampling required for the completion of a random tensor is in general close to $\mathcal{O}(nrlog(n))$ as in the matrix case [7].

current values in G.
As results, we plot the number of successful recoveries ($R_C/\|M_C\| < 10^{-5}$) for different mode sizes n (each single tuple uniform), dimensions d and maximal ranks k of the target tensor (Figures 5, 6).

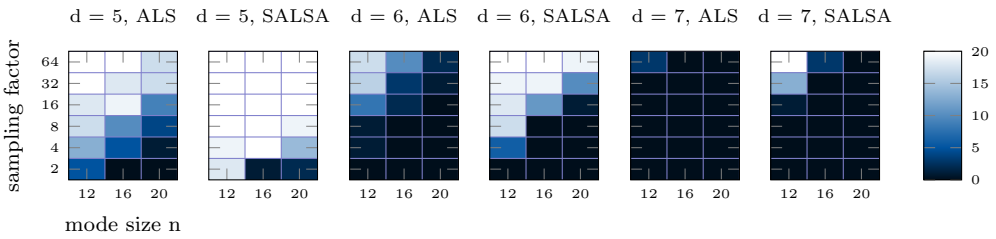

Figure 5: ($d = 5, 6, 7$, $r_P = 6$, $r_{lim} = 9$, $n = 8, 12, 16, 20$, $C_{sf} = 2, 4, 8, 16, 32, 64$) Displayed as 20 shades of blue (black (0) to white (all 20)) are the number of successful reconstructions for random tensors with maximal rank $k = 6$ for ALS and SALSA

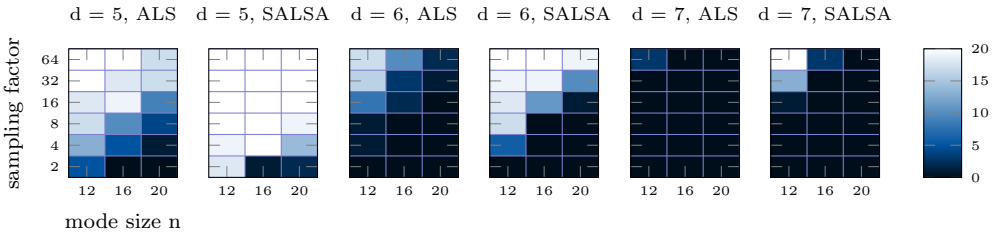

Figure 6: ($d = 5, 6, 7$, $r_P = 8$, $r_{lim} = 11$, $n = 12, 16, 20$, $C_{sf} = 2, 4, 8, 16, 32, 64$) Displayed as 20 shades of blue (black (0) to white (all 20)) are the number of successful reconstructions for random tensors with maximal rank $k = 8$ for ALS and SALSA

7.7 Recovery of the Rank Adaption Test Tensor

Last but not least, we consider the recovery of tensors as in Example 1.5, 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 Figure 7, we refer to Section 7.6.

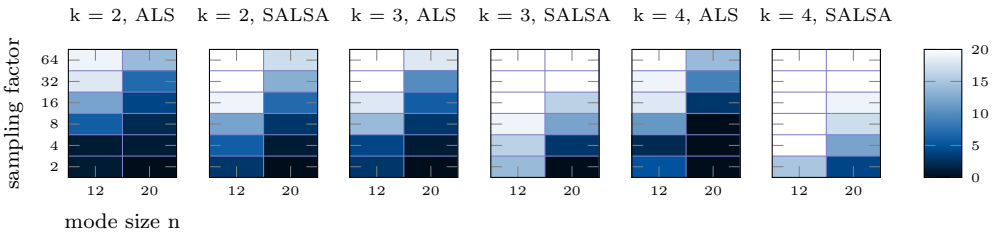

Figure 7: ($d = 6$, $r_P = 2k$, $r_{lim} = 2k + 3$, $n = 12, 20$, $C_{sf} = 2, 4, 8, 16, 32, 64$) Displayed as 20 shades of blue (black (0) to white (all 20)) are the number of successful reconstructions for the rank adaption test tensor with rank $(1, k, k, k, 1, 2k, 1)$ for ALS and SALSA

7.8 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, but SALSA yields better results (the timing however is worse). The two generic functions show that the residuals can be multiple orders of magnitude better, and although the functions were chosen quite randomly, we do not want to over-interprete 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 mentioned before, the tuning parameters of SALSA could be relaxed to better keep up with the speed of ALS in case of larger sampling.

8 Conclusions

In this article, we have demonstrated that the most successful completion algorithms do not behave continuously with 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 micro-steps in order to uncouple the optimization of a discrete, technical rank. While the exact derivation and implementation of SALSA is presumably improvable, we take the notable numerical results as indication that *stability (under truncation)* is a worthwhile property. Briefly said, SALSA can crack harder problems by investing an advanced amount of time. Let it be mentioned that, although we focused on tensor completion (with possibly small sampling sets), the derivations given in this paper allow for a straightforward generalization to arbitrary semi-elliptic problems.
The computational complexity remains the same and it poses an open question whether it can be reduced. Furthermore, it may be possible to adapt the presented ideas to manifold based method.

9 Appendix A (Proofs)

Construction of the tensor in Example 1.5:
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.

Lemma 4.5:

Proof. Let $V = V_i^+ \cup V_i^-$, $V_i^{+/-} := \{X \in \mathbb{R}^{n \times m} \mid X_{i1} \gtrless 0,\ \|X\|_F = \omega\}$. We can split the integral and simplify

$$\begin{aligned}
Y := (\int_{V_i^+} + \int_{V_i^-}) X^T H X \,\mathrm{d}X = \int_{V_i^+} &(X_{<i,-} \mid X_{i,-} \mid X_{>i,-})^T H (X_{<i,-} \mid X_{i,-} \mid X_{>i,-})\\
&+ (X_{<i,-} \mid -X_{i,-} \mid X_{>i,-})^T H (X_{<i,-} \mid -X_{i,-}, \mid X_{>i,-}) \,\mathrm{d}X
\end{aligned}$$

Hence, for $i \neq j$,

$$Y_{ij} = \int_{V_i^+} X_{i,-}^T H X_{j,-} + (-X_{i,-})^T H X_{j,-} \,\mathrm{d}X = 0.$$

It follows that the matrix Y is diagonal and must therefore, considering permutations P, s.t. $V = PV$, be a multiple of I_m. Now, let $H + H^T = Q^T D Q$ be an eigenvalue decomposition. Then $\mathrm{tr}(D) = 2\mathrm{tr}(H)$ and since Q is orthogonal, we have

$$2\mathrm{tr}(Y) = \int_V \mathrm{tr}(X^T D X) \,\mathrm{d}X = \sum_{i=1}^n \int_V \mathrm{tr}(X_{i,-}^T d_i X_{i,-}) \,\mathrm{d}X = \sum_{i=1}^n d_i \int_V X_{i,-} X_{i,-}^T \,\mathrm{d}X$$

Further, due to symmetry

$$n \int_V X_{j,-} X_{j,-}^T \,\mathrm{d}X = \int_V \sum_{i=1}^n X_{i,-} X_{i,-}^T \,\mathrm{d}X = \int_V \mathrm{tr}(X^T X) \,\mathrm{d}X$$

for any j. Thereby

$$\mathrm{tr}(Y) = \frac{1}{2n} \sum_{i=1}^n d_i \int_V \mathrm{tr}(X^T X) \,\mathrm{d}X = \frac{\mathrm{tr}(H)}{n} \int_V \mathrm{tr}(X^T X) \,\mathrm{d}X = \omega^2 \frac{\mathrm{tr}(H)}{n} |V|.$$

This then gives the result. □

Theorem 4.7:

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

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

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

$$C_{B,S,\mathcal{R},\Gamma\mathcal{N}\Theta,\mathcal{L}}(\widetilde{N}) \propto \int_{\mathbb{V}_\omega} \|(\mathcal{L}+\Delta\mathcal{L}\Gamma^{-1})\widetilde{N}(\mathcal{R}+\Theta^{-1}\Delta\mathcal{R}) - B\|_S^2 \,\mathrm{d}\Delta\mathcal{L}\mathrm{d}\Delta\mathcal{N}\mathrm{d}\Delta\mathcal{R},$$
$$\mathbb{V}_\omega = \{(\Delta\mathcal{L}, \Delta\mathcal{N}, \Delta\mathcal{R}) \mid \|\Delta\mathcal{L}\|^2 + \|\Delta\mathcal{N}\|^2 + \|\Delta\mathcal{R}\|^2 \leq \omega^2\}. \tag{9.1}$$

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

$$N^+(j) = \operatorname*{argmin}_{\widetilde{N}(j)} \int_{\mathbb{V}_\omega} \|((\mathcal{R}+\Theta^{-1}\Delta\mathcal{R})^T \otimes_K (\mathcal{L}+\Delta\mathcal{L}\Gamma^{-1}))\ \mathrm{vec}(\widetilde{N}(j)) \tag{9.2}$$
$$- \mathrm{vec}(B(j))\|^2_{\mathrm{vec}(S(j))} \,\mathrm{d}\Delta\mathcal{L}\mathrm{d}\Delta\mathcal{N}\mathrm{d}\Delta\mathcal{R} \tag{9.3}$$

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 $H(j) \in \mathbb{R}^{|\mathcal{I}|/n_N \times |\mathcal{I}|/n_N}$ 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

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

and

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

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

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

Now, let $\ell = \#_{\mathcal{R}},\ n = \#_{\mathcal{N}},\ k = \#_{\mathcal{L}}$. 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

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

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

$$\int_{\mathbb{V}} f \,\mathrm{d}\delta = \int_{\lambda=0}^{\omega} \int_{S_\lambda^{n+\ell+k-1}} f \,\mathrm{d}\delta\mathrm{d}\lambda = \int_{\lambda=0}^{\omega} \lambda \int_{g=0}^{\pi/2} \int_{S^{n-1}_{\lambda\sin(g)}} \int_{S^{\ell+k-1}_{\lambda\cos(g)}} f \,\mathrm{d}\delta\mathrm{d}g\mathrm{d}\lambda =$$

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

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

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

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 4.6 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 $\mathcal{R}^T$ ($\delta_1 = 0$) or $\Delta\mathcal{R}^T\Theta^{-1}$ ($\delta_1 = 1$) and analogously Y as $\mathcal{L}$ ($\delta_2 = 0$) or $\Delta\mathcal{L}\Gamma^{-1}$ ($\delta_2 = 1$). 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

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

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:

$$\widetilde{Y}(0,0) = C_H(0,0),\ \widetilde{Y}(1,0) = \zeta_1 C_H(1,0),\ \widetilde{Y}(0,1) = \zeta_2 C_H(0,1),\ \widetilde{Y}(1,1) = \zeta_{1,2} C_H(1,1),$$

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

Lemma 5.1:

Proof. First,

$$\zeta_1^{(\mu)} = \omega^2 s_1^2 \frac{\prod_{s=1}^{\mu-1} n_s}{E} = \omega^2 \frac{\sum_{s=1}^{\mu-1} n_s}{\sum_{s=1}^{d} n_s},$$

with an analog result for $\zeta_2^{(\mu)}$. For the mixed term, we have

$$\zeta_{1,2}^{(\mu)} = \zeta_1^{(\mu)}\zeta_2^{(\mu)}\frac{E}{E+2} = \zeta_1^{(\mu)}\zeta_2^{(\mu)}(1 - \frac{2}{E+2}).$$

□

10 Appendix B (Behavior of the SALSA Filter)

We investigate the behavior of the filter $\mathcal{F}$ as defined by (4.8) and its relevance for SALSA in order to motivate Definitions 6.1, 6.2 and 6.9. Throughout this section, we assume that the sampling is such that the minimizer in Theorem 4.7 is basically equal to

$$N^+ = \mathcal{F} \odot (\mathcal{L}^T\ B\ \mathcal{R}^T), \tag{10.1}$$

which at last holds for $P = \mathcal{I}$ (cf. Corollary 4.8). Since $\zeta_1\zeta_2 = \zeta_{1,2}$ (cf. (5.1)), we can rewrite

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

We are interested in the fixpoints of this update, i.e. we postulate $N^+ = \Gamma\ \mathcal{N}\ \Theta$. Then, since $\mathfrak{R}(\mathcal{N}\ \Theta)$ is row-orthogonal (cf. Lemma 4.2), it holds

$$\begin{aligned} D_{\zeta_1}(\Gamma)\ Z &= \Gamma, \\ Z &= \mathfrak{R}((\mathcal{L}^T\ B\ \mathcal{R}^T)\ D_{\zeta_2}(\Theta))\ \mathfrak{R}(\mathcal{N}\ \Theta)^T, \end{aligned} \tag{10.2}$$

where $Z =: \mathrm{diag}(\sigma^{(Z)})$ is necessarily a diagonal matrix (certainly, an analogous argument holds for Θ as well). The focus of our analysis is hence on the fixpoints of the function $d_{\sigma^{(Z)},c} : \sigma \mapsto (1 + c\sigma^{-2})^{-1}\sigma^{(Z)}$, because (10.2) can only hold if $d_{\sigma^{(Z)},\zeta_1}(\gamma_i) = \gamma_i$ for all i. For each pair $(\sigma^{(Z)}, c)$, the only attractive fixpoint (if existent) is given by $f_{\mathrm{stab}} = \frac{1}{2}\sigma^{(Z)} + \frac{1}{2}\sqrt{(\sigma^{(Z)})^2 - 4c}$ and the repelling one by $f_{\mathrm{rep}} = \frac{1}{2}\sigma^{(Z)} - \frac{1}{2}\sqrt{(\sigma^{(Z)})^2 - 4c}$. At the point where $f_{\mathrm{stab}} = f_{\mathrm{rep}}$, it holds $\sigma = c = \frac{1}{2}\sigma^{(Z)}$. 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 Figure 8. The relation to the filter is given by

$$\mathcal{F}_{i,1} = (D_{\zeta_1}(\Gamma))_{i,i} \cdot \underbrace{(D_{\zeta_2}(\Theta))_{1,1}}_{\approx 1}.$$

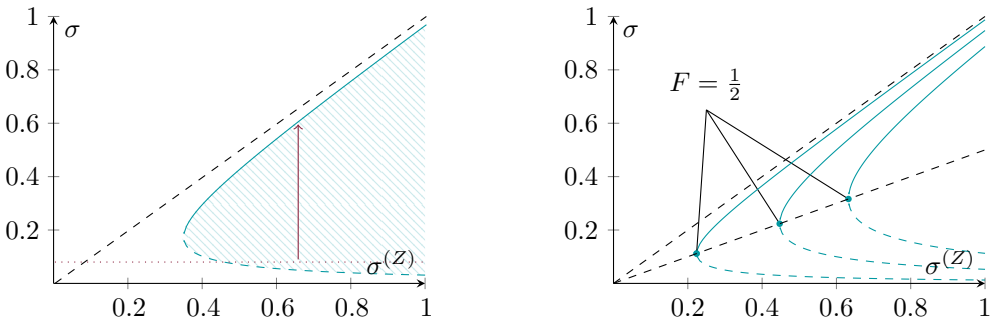

Figure 8: Left: Plotted are the fixpoints (continuous for attractive, dashed for repelling ones, in teal) of $d_{\sigma^{(Z)},c}$ for one fixed c with repect to $\sigma^{(Z)}$. Within the hatched area, singular values rise until they reach the upper boundary. A lower limit to the singular values is indicated 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 $1/2$.

A (stabilized) singular value corresponds to some attractive fixpoint of $d_{\sigma^{(Z)},c}$. Therefore it necessarily holds $(D_{\zeta_1}(\Gamma))_{i,i} > 0.5$. In practice, the value F_{stab} should be chosen larger, as well as F_{virt} lower, not only to reduce the computational cost, but also to avoid premature reactions within the optimization. Since the singular values Γ take part in another, neighboring micro-step as well, the accordant value is also taken into account (cf. 4.8).
It is now easy to understand why a lower limit to all singular values is required. As displayed in Figure 8 (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 an accordant micro-step. So we cannot allow it to converge to zero.

References

[1] Bachmayr, M., Schneider, R.: Iterative methods based on soft thresholding of hierarchical tensors. Foundations of Computational Mathematics pp. 1–47 (2016). DOI 10.1007/s10208-016-9314-z. URL http://dx.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. Foundations of Computational Mathematics pp. 1–50 (2016). DOI 10.1007/s10208-016-9317-9. URL http://dx.doi.org/10.1007/s10208-016-9317-9

[3] Ballani, J., Grasedyck, L.: A projection method to solve linear systems in tensor format. Numerical Linear Algebra with Applications **20**(1), 27–43 (2013). DOI 10.1002/nla.1818. URL http://dx.doi.org/10.1002/nla.1818

[4] Ballani, J., Grasedyck, L., Kluge, M.: Black box approximation of tensors in hierarchical tucker format. Linear Algebra and its Applications **438**(2), 639 – 657 (2013). DOI http://dx.doi.org/10.1016/j.laa.2011.08.010. URL http://www.sciencedirect.com/science/article/pii/S002437951100591X

[5] Beylkin G., M.M.: Numerical operator calculus in higher dimensions. PNAS **99**(16), 10,246–10,251 (2002). DOI 10.1073/pnas.112329799

[6] Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Foundations of Computational Mathematics **9**(6), 717 (2009). DOI 10.1007/s10208-009-9045-5. URL http://dx.doi.org/10.1007/s10208-009-9045-5

[7] Candès, E.J., Tao, T.: The power of convex relaxation: Near-optimal matrix completion. IEEE Trans. Inf. Theor. **56**(5), 2053–2080 (2010). DOI 10.1109/TIT.2010.2044061. URL http://dx.doi.org/10.1109/TIT.2010.2044061

[8] Dolgov, S.V., Savostyanov, D.V.: Alternating minimal energy methods for linear systems in higher dimensions. SIAM Journal on Scientific Computing **36**(5), A2248–A2271 (2014). DOI 10.1137/140953289. URL http://dx.doi.org/10.1137/140953289

[9] Dopico, F.M.: A note on sin θ theorems for singular subspace variations. BIT Numerical Mathematics **40**(2), 395–403 (2000). DOI 10.1023/A:1022303426500. URL http://dx.doi.org/10.1023/A:1022303426500

[10] Espig, M., Khachatryan, A.: Convergence of alternating least squares optimisation for rank-one approximation to high order tensors. arXiv:1503.05431 (2015). URL https://arxiv.org/abs/1503.05431

[11] Gandy, S., Recht, B., Yamada, I.: Tensor completion and low-n-rank tensor recovery via convex optimization. Inverse Problems **27**(2), 025,010 (2011). URL http://stacks.iop.org/0266-5611/27/i=2/a=025010

[12] Grasedyck, L.: Hierarchical singular value decomposition of tensors. SIAM Journal on Matrix Analysis and Applications **31**(4), 2029–2054 (2010). DOI 10.1137/090764189. URL http://dx.doi.org/10.1137/090764189

[13] Grasedyck, L., Kluge, M., Krämer, S.: Variants of alternating least squares tensor completion in the tensor train format. SIAM Journal on Scientific Computing **37**(5), A2424–A2450 (2015). DOI 10.1137/130942401. URL http://dx.doi.org/10.1137/130942401

[14] Grasedyck, L., Kressner, D., Tobler, C.: A literature survey of low-rank tensor approximation techniques. GAMM-Mitteilungen **36**(1), 53–78 (2013). DOI 10.1002/gamm.201310004. URL http://dx.doi.org/10.1002/gamm.201310004

[15] Gross, D.: Recovering low-rank matrices from few coefficients in any basis. IEEE Transactions on Information Theory **57**(3), 1548–1566 (2011). DOI 10.1109/TIT.2011.2104999

[16] Hackbusch, W.: Numerical tensor calculus. Acta Numerica **23**, 651–742 (2014). DOI 10.1017/S0962492914000087. URL https://www.cambridge.org/core/article/numerical-tensor-calculus/67876F5C81E4D4F84CA334E204B6EADC

[17] Hackbusch, W., Kühn, S.: A new scheme for the tensor representation. Journal of Fourier Analysis and Applications **15**(5), 706–722 (2009). DOI 10.1007/s00041-009-9094-9. URL http://dx.doi.org/10.1007/s00041-009-9094-9

[18] Hackbusch, W., Schneider, R.: Tensor Spaces and Hierarchical Tensor Representations, pp. 237–261. Springer International Publishing, Cham (2014). DOI 10.1007/978-3-319-08159-5_12. URL http://dx.doi.org/10.1007/978-3-319-08159-5_12

[19] Holtz, S., Rohwedder, T., Schneider, R.: The alternating linear scheme for tensor optimization in the tensor train format. SIAM Journal on Scientific Computing **34**(2), A683–A713 (2012). DOI 10.1137/100818893. URL http://dx.doi.org/10.1137/100818893

[20] Holtz, S., Rohwedder, T., Schneider, R.: On manifolds of tensors of fixed tt-rank. Numerische Mathematik **120**(4), 701–731 (2012). DOI 10.1007/s00211-011-0419-7. URL http://dx.doi.org/10.1007/s00211-011-0419-7

[21] Jeckelmann, E.: Dynamical density-matrix renormalization-group method. Phys. Rev. B **66**, 045,114 (2002). DOI 10.1103/PhysRevB.66.045114. URL http://link.aps.org/doi/10.1103/PhysRevB.66.045114

[22] Krämer, S.: The geometrical description of feasible singular values in the tensor train format. (in preparation, title may yet change)

[23] Kressner, D., Steinlechner, M., Vandereycken, B.: Low-rank tensor completion by riemannian optimization. BIT Numerical Mathematics **54**(2), 447–468 (2014). DOI 10.1007/s10543-013-0455-z. URL http://dx.doi.org/10.1007/s10543-013-0455-z

[24] Liu, Y., Shang, F.: An efficient matrix factorization method for tensor completion. IEEE Signal Processing Letters **20**(4), 307–310 (2013). DOI 10.1109/LSP.2013.2245416

[25] Matthies, H.G., Zander, E.: Solving stochastic systems with low-rank tensor compression. Linear Algebra and its Applications **436**(10), 3819 – 3838 (2012). DOI http://dx.doi.org/10.1016/j.laa.2011.04.017. URL http://www.sciencedirect.com/science/article/pii/S0024379511003223

[26] Mu, C., Huang, B., Wright, J., Goldfarb, D.: Square deal: Lower bounds and improved relaxations for tensor recovery. In: T. Jebara, E.P. Xing (eds.) Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 73–81. JMLR Workshop and Conference Proceedings (2014). URL http://jmlr.org/proceedings/papers/v32/mu14.pdf

[27] Oseledets, I., Tyrtyshnikov, E.: Tt-cross approximation for multidimensional arrays. Linear Algebra and its Applications **432**(1), 70 – 88 (2010). DOI http://dx.doi.org/10.1016/j.laa.2009.07.024. URL http://www.sciencedirect.com/science/article/pii/S0024379509003747

[28] Oseledets, I.V.: Tensor-train decomposition. SIAM Journal on Scientific Computing **33**(5), 2295–2317 (2011). DOI 10.1137/090752286. URL http://dx.doi.org/10.1137/090752286

[29] Oseledets, I.V., Tyrtyshnikov, E.E.: Breaking the curse of dimensionality, or how to use svd in many dimensions. SIAM Journal on Scientific Computing **31**(5), 3744–3759 (2009). DOI 10.1137/090748330. URL http://dx.doi.org/10.1137/090748330

[30] Rauhut, H., Schneider, R., Stojanac, Ž.: Tensor Completion in Hierarchical Tensor Representations, pp. 419–450. Springer International Publishing, Cham (2015). DOI 10.1007/978-3-319-16042-9_14. URL http://dx.doi.org/10.1007/978-3-319-16042-9_14

[31] Recht, B.: A simpler approach to matrix completion. Journal of Machine Learning Research **12**, 3413–3430 (2011)

[32] Rohwedder, T., Uschmajew, A.: On local convergence of alternating schemes for optimization of convex problems in the tensor train format. SIAM Journal on Numerical Analysis **51**(2), 1134–1162 (2013). DOI 10.1137/110857520. URL http://dx.doi.org/10.1137/110857520

[33] Signoretto, M., TranDinh, Q., De Lathauwer, L., Suykens, J.A.K.: Learning with tensors: a framework based on convex optimization and spectral regularization. Machine Learning **94**(3), 303–351 (2014). DOI 10.1007/s10994-013-5366-3. URL http://dx.doi.org/10.1007/s10994-013-5366-3

[34] Silva, C.D., Herrmann, F.J.: Optimization on the hierarchical tucker manifold applications to tensor completion. Linear Algebra and its Applications **481**, 131 – 173 (2015). DOI http://dx.doi.org/10.1016/j.laa.2015.04.015. URL http://www.sciencedirect.com/science/article/pii/S0024379515002530

[35] Vidal, G.: Efficient classical simulation of slightly entangled quantum computations. Phys. Rev. Lett. **91**, 147,902 (2003). DOI 10.1103/PhysRevLett.91.147902. URL http://link.aps.org/doi/10.1103/PhysRevLett.91.147902

[36] Wedin, P.Å.: Perturbation bounds in connection with singular value decomposition. BIT Numerical Mathematics **12**(1), 99–111 (1972). DOI 10.1007/BF01932678. URL http://dx.doi.org/10.1007/BF01932678

[37] Wen, Z., Yin, W., Zhang, Y.: Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. Mathematical Programming Computation **4**(4), 333–361 (2012). DOI 10.1007/s12532-012-0044-1. URL http://dx.doi.org/10.1007/s12532-012-0044-1

[38] Weyl, H.: Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). Mathematische Annalen **71**(4), 441–479 (1912). DOI 10.1007/BF01456804. URL http://dx.doi.org/10.1007/BF01456804

[39] White, S.R.: Density matrix formulation for quantum renormalization groups. Phys. Rev. Lett. **69**, 2863–2866 (1992). DOI 10.1103/PhysRevLett.69.2863. URL http://link.aps.org/doi/10.1103/PhysRevLett.69.2863

11 Appendix C (Experimental Data)

Following are the precise values for Figures 3 and 4:

$n = 12$		ALS			SALSA		
d	C_{sf}	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$
6	2	6.3e-04(1.9)	8.4e-05(2.5)	90(31)	3.1e-04(1.1)	2.1e-05(1.9)	391(78)
	4	2.2e-04(1.8)	6.5e-05(2.3)	78(17)	1.3e-04(1.2)	2.8e-05(1.3)	615(62)
	6	1.4e-04(1.1)	5.5e-05(1.1)	86(7)	8.8e-05(1.1)	3.2e-05(1.1)	892(52)
9	2	7.5e-04(2.2)	2.3e-04(4.2)	145(71)	1.7e-04(1.1)	1.1e-05(1.6)	1031(146)
	4	8.0e-05(1.1)	2.0e-05(1.1)	306(30)	4.0e-05(1.1)	9.8e-06(1.1)	2370(201)
	6	6.5e-05(1.1)	2.5e-05(1.2)	295(12)	3.1e-05(1.1)	1.2e-05(1.0)	3291(423)
15	2	6.6e-04(1.5)	3.3e-04(2.3)	444(261)	5.6e-04(1.6)	1.7e-04(3.0)	2132(633)
	4	3.2e-05(1.1)	6.9e-06(1.1)	1219(60)	3.2e-05(1.3)	5.1e-06(1.4)	11498(1886)
	6	2.7e-05(1.1)	8.7e-06(1.1)	1575(78)	1.1e-05(1.4)	3.4e-06(1.3)	20186(4826)

$n = 20$		ALS			SALSA		
d	C_{sf}	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$
6	2	3.8e-03(1.1)	1.8e-03(1.1)	79(40)	9.2e-04(1.2)	6.6e-05(2.3)	632(141)
	4	5.5e-04(1.2)	1.7e-04(1.2)	155(46)	3.7e-04(1.2)	5.3e-05(1.2)	918(136)
	6	4.6e-04(1.3)	1.5e-04(1.5)	224(69)	2.9e-04(1.2)	6.7e-05(1.1)	1271(129)
9	2	2.6e-03(1.0)	1.4e-03(1.1)	170(72)	5.3e-04(1.1)	3.2e-05(1.2)	1334(266)
	4	2.9e-04(1.1)	1.0e-04(1.1)	363(31)	2.4e-04(1.1)	4.5e-05(1.3)	3119(629)
	6	2.0e-04(1.1)	7.7e-05(1.2)	677(123)	1.3e-04(1.1)	4.4e-05(1.1)	5104(1302)
15	2	1.6e-03(1.0)	8.6e-04(1.0)	638(217)	1.7e-03(1.1)	7.1e-04(1.2)	3152(491)
	4	1.1e-04(1.1)	2.4e-05(1.2)	1994(144)	1.1e-04(1.1)	1.9e-05(1.1)	15622(2465)
	6	8.2e-05(1.0)	2.8e-05(1.1)	2929(91)	4.9e-05(1.5)	1.6e-05(1.4)	32167(9352)

Table 1: Results for Subsection 7.4 (with arithmetic and geometric variances in brackets) using a (pure) Matlab implementation

$d = 8, n = 8$	ALS			SALSA		
C_{sf}	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$
2	6.7e-03(3.3)	3.2e-03(6.2)	82(59)	2.4e-04(10.0)	9.3e-06(36.5)	498(179)
4	4.2e-03(1.2)	3.8e-03(1.3)	38(22)	1.1e-05(3.8)	1.7e-07(15.4)	1276(523)
6	3.8e-03(1.2)	3.6e-03(1.2)	40(25)	2.1e-06(4.6)	4.9e-08(11.7)	2278(967)

$d = 7, n = 8$	ALS			SALSA		
C_{sf}	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$
2	1.4e-01(6.4)	3.1e-02(8.9)	76(98)	1.9e-03(2.4)	3.0e-05(3.5)	403(93)
4	6.5e-04(39.5)	2.2e-04(63.8)	99(59)	2.9e-05(1.5)	6.9e-06(1.0)	954(627)
6	6.9e-05(23.3)	4.2e-05(26.8)	84(32)	1.5e-05(1.1)	8.2e-06(1.0)	659(45)

$d = 11, n = 8$	ALS			SALSA		
C_{sf}	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$	$\langle R_C/\|M_C\|\rangle_{\text{geo}}$	$\langle R_P/\|M_P\|\rangle_{\text{geo}}$	$\langle\text{time}\rangle_{\text{ar}}$
2	4.2e-02(1.7)	3.8e-02(2.3)	52(43)	2.0e-03(6.4)	4.6e-04(16.8)	759(493)
4	3.1e-02(3.4)	2.3e-02(6.6)	110(134)	6.4e-05(1.3)	9.7e-06(1.6)	2527(752)
6	3.4e-02(4.3)	3.0e-02(6.2)	99(152)	3.8e-05(1.4)	7.5e-06(1.8)	4657(1162)

Table 2: Results for Subsection 7.5 (with arithmetic and geometric variances brackets) using a (pure) Matlab implementation