ASYMPTOTIC LOG-DET SUM-OF-RANKS MINIMIZATION VIA TENSOR (ALTERNATING) ITERATIVELY REWEIGHTED LEAST SQUARES

SEBASTIAN KRAMER

Abstract. Affine sum-of-ranks minimization (ASRM) generalizes the affine rank minimization (ARM) problem from matrices to tensors. Here, the interest lies in the ranks of a family \( K \) of different matricizations. Transferring our priorly discussed results on asymptotic log-det rank minimization, we show that iteratively reweighted least squares with weight strength \( p = 0 \) remains a, theoretically and practically, particularly viable method denoted as IRLS-0\( K \). As in the matrix case, we prove global convergence of asymptotic minimizers of the log-det sum-of-ranks function to desired solutions. Further, we show local convergence of IRLS-0\( K \) in dependence of the rate of decline of the therein appearing regularization parameter \( \gamma \downarrow 0 \). For hierarchical families \( K \), we show how an alternating version (AIRLS-0\( K \), related to prior work under the name SALSA) can be evaluated solely through tensor tree network based operations. The method can thereby be applied to high dimensions through the avoidance of exponential computational complexity. Further, the otherwise crucial rank adaption process becomes essentially superfluous even for completion problems. In numerical experiments, we show that the therefor required subspace restrictions and relaxation of the affine constraint cause only a marginal loss of approximation quality. On the other hand, we demonstrate that IRLS-0\( K \) allows to observe the theoretical phase transition also for generic tensor recoverability in practice. Concludingly, we apply AIRLS-0\( K \) to larger scale problems.

Key words. affine rank minimization, iteratively reweighted least square, matrix recovery, matrix completion, log-det function

AMS subject classifications. 15A03, 15A29, 65J20, 90C31, 90C26

1. Introduction

The setting of affine sum-of-ranks minimization (ASRM) is a generalization of the affine rank minimization (ARM) problem for matrices to tensors. While the tensor rank refers to the minimal number of elementary tensors required for a decomposition into a sum, we are here interested in the ranks of so called matricizations. Let \( [d] = \{1, \ldots, d\} \), \( d \in \mathbb{N} \), as well as \( n_\mu \in \mathbb{N}, \mu = 1, \ldots, d \). For \( \emptyset \neq J \subseteq [d] \) and \( J^c := [d] \setminus J \), we define such matricizations (cf. [15])

\[
\left(v \right)[J] : \mathbb{R}^{n_1} \otimes \cdots \otimes \mathbb{R}^{n_d} \to \mathbb{R}^{n_J \times n_{J^c}},
\]

\[n_S := \prod_{\mu \in S} n_\mu,\]

as the simple reshaping isomorphisms induced via

\[
(v_1 \otimes \cdots \otimes v_d)[J] := \text{vec}(\bigotimes_{j \in J} v_j) \cdot \text{vec}(\bigotimes_{j \in J^c} v_j)^T, \quad v_i \in \mathbb{R}^{n_i}, \quad i = 1, \ldots, d,
\]
where \( \text{vec}(\cdot) : \mathbb{R}^{n_d \times n_d} \rightarrow \mathbb{R}^{n_1 \times \cdots \times n_d} \) denotes the vectorization in co-lexicographic (column-wise) order. As usual, we identify \( \mathbb{R}^{n_1 \otimes \cdots \otimes R^{n_d}} \cong \mathbb{R}^{n_1 \times \cdots \times n_d} \). For a (not necessarily hierarchical) family of subsets \( \mathcal{K} \subseteq \{ J \subseteq [d] \mid J \neq \emptyset \} \) and a surjective linear operator \( \mathcal{L} : \mathbb{R}^{n_1 \times \cdots \times n_d} \rightarrow \mathbb{R}^\ell, \ell < n_{[d]}, \) as well as measurements \( y \in \text{image}(\mathcal{L}) \), we then define ASRM to refer to the problem of finding

\[
\arg\min_{X \in \mathbb{R}^{n_1 \times \cdots \times n_d}} \sum_{J \in \mathcal{K}} \text{rank}(X[J]) \quad \text{subject to} \quad \mathcal{L}(X) = y.
\]  

(1.1)

This setting is not only of particular interest due to its regularizing properties, but its close relation to so-called hierarchical (or tensor tree) decompositions (cf. [15, 25]). We are here however mainly interested in the problem itself, and only secondarily in the possibility to recover an eventual ground truth tensor from its measurements. As large parts of this work rely on our preceding article [26], which in turn is based on [5, 7, 11, 29], we strongly recommend to take notice of such.

1.1. Approaches to ASRM and tensor recovery. Affine rank minimization (ARM), as theoretical origin of ASRM, is included as such for the dimension \( d = 2 \) and \( \mathcal{K} = \{ \{1\} \} \) [26], and consequently defined as the problem to find a matrix

\[
X^* \in \arg\min_{X \in \mathbb{R}^{n_1 \times n_2}} \text{rank}(X) \quad \text{subject to} \quad \mathcal{L}(X) = y.
\]

This setting in turn is based on the affine cardinality minimization problem (ACM), that is to find a vector

\[
x^* \in \arg\min_{x \in \mathbb{R}^n} \text{card}(x) \quad \text{subject to} \quad \mathcal{L}(x) = y.
\]

A short overview over recovery methods as well as the role of iteratively reweighted least squares (IRLS, cf. [11, 29] for ARM and [5, 7] for ACM) for these two problems can be found in our preceding article [26]. To the best of our knowledge, IRLS has only priorly been considered with regard to the ASRM problem for tensors in the thesis [25], from which also the related, so-called stable ALS approximation algorithm [16] stems. Relaxations of ASRM itself however have been considered before, including the minimization of the sum of nuclear norms [12, 28, 35]. The tensor rank as outlined in the introduction, Section 1, however, is hard to calculate, and usually not the direct target of minimization. Though [37, 38] utilize the canonical polyadic decomposition to a certain fiber completion problem. Other algorithms rely on the explicit, separate adaption of unknown ranks such as low rank manifold [27, 36, 39] or a-priori representation or subspace based optimization [18, 19, 34]. However, non-intrusive rank adaption schemes, even if elaborate, tend to be problematic [16]. The AIRLS related method presented therein, as well as [2, 13, 14] based thereon, contrarily consider an intrusive regularization related to reweighting that circumvents the instability and overfitting problems otherwise caused. Another class of algorithm requires to choose specific sampling points, prominently cross approximation based methods [1, 22, 31]. Though such are preferable in that setting, we here however assume the affine measurement operator to be a priori given.

1.2. Contributions and organization of this paper. The novel aspects of this paper are organized as follows.

- In Sections 1.3 and 1.4, we generalize the optimization as well as reweighting process from the matrix to the tensor case in an introductory manner. Section 1.5 contains a preliminary description of hierarchical decompositions and the thereto related data sparse optimization.

- In Section 2, we interpret the tensor log-det approach as successive minimization scheme and thereby prove the convergence of global optima to the desired solution, as analogously done for the matrix case [26].
1.3. Asymptotic minimization. We have priorly discussed in [26] as based on [5,7,11,29] in which way the ARM problem for matrices can be approached via the asymptotic minimization (cf. Definition 1.1) of the family

\[ f_\gamma(A) := \log \prod_{i=1}^{k_1} (\sigma_i^2(A) + \gamma) = \log \det(AA^T + \gamma I), \quad \gamma \searrow 0, \]

for which \( \sigma_i(A), \ i = 1, \ldots, r, \) are defined as the singular values of \( A \in \mathbb{R}^{k_1 \times k_2} \) and \( \sigma_i(A) = 0, \ i > r, \ r = \text{rank}(A). \) Plainly analogous, its tensor version for the minimization of a sum of ranks is defined as (see Section 2)

\[ f^K_\gamma(X) := \sum_{J \in K} f_\gamma(X^{[J]}) = \log \prod_{J \in K} \prod_{i=1}^{n_J} (\sigma_i^{(J)}(X)^2 + \gamma), \]

where \( \sigma_i^{(J)}(X) = \sigma_i(X^{[J]}) \) is the \( i \)-th singular value of the matrix \( X^{[J]} \in \mathbb{R}^{n_J \times n_J}. \) Thus the matrix version corresponds to \( K = \{\{1\}\} \), whereas for the alternating IRLS method, we have also considered the complementary \( K = \{\{2\}\}. \) In [26], we have already reasoned the choice \( p = 0 \) of the therein appearing weight strength parameter \( p \in [0, 1]. \) Thus, we here only regard the thereto corresponding log-det approach laid out above, as opposed to the other extreme \( p = 1 \) associated to nuclear norm minimization. This leads us to the following, potential solutions to the ASRM problem.

\[ \text{Most formulas are however easily adaptable to } p \in [0, 1]. \]
Definition 1.1. We define

\[ X^*: = \{ X^* \mid \exists (X_\gamma)_{\gamma \geq 0} \subset L^{-1}(y), \ X^* = \lim_{\gamma \to 0} X_\gamma, \ f_\gamma^K(X_\gamma) = \min_{X \in L^{-1}(y)} f_\gamma^K(X) \}. \]

This set of asymptotic, global minimizers indeed yields the desired solutions as we prove in Theorem 2.4. The decline of the parameter \( \gamma \) is no less important here as more detailly remarked on in the predecessor [26]. It should further be noted that neither the ranks \( r(J) := \text{rank}(X^{[J]}) \) (cf. Section 2), nor the families of singular values \( \sigma^{(J)}, J \in \mathcal{K} \), are independent of each other [24], though not prohibitively so in regard of aboves approach.

1.4. Iteratively reweighted least squares (IRLS). In line with the overall generalization, also iteratively reweighted least squares (IRLS) allows to be applied to the minimization of a sum of ranks of a tensor. For the matrix case, one version (cf. [26, 29]) defines (\( \| \cdot \|_F \) being the Frobenius norm)

\[ X^{(i)} := \arg\min_{X \in L^{-1}(y)} \| W^{1/2}_{\gamma,1} X \|_F, \quad W_{\gamma,X} := (XX^T + \gamma I)^{-1}, \]

for a monotonically decreasing sequence \( \{ \gamma^{(i)} \}_{i \geq 0} \subset \mathbb{R}_{>0} \). The tensor variant straightforwardly is given by (see Section 3)

\[ \gamma^{(i)} := \arg\min_{X \in L^{-1}(y)} \sum_{J \in \mathcal{K}} \| (W^{(J)}_{\gamma,1}, X^{[1]} - 1)^{1/2} (X^{(i)} - 1) \|_F^2, \]

where the weight matrices follow the same generalization with

\[ W^{(J)}_{\gamma,X} := W_{\gamma,X^{[J]}} = (X^{[J]}(X^{[J]})^T + \gamma I)^{-1}, \quad J \in \mathcal{K}. \]

Continued from the vector as well as matrix case, it also here holds true that for a sequence \( X_\gamma \to \mathbf{X} \) with sufficiently fast declining singular values

\[ \sum_{J \in \mathcal{K}} \| W^{(J)}_{\gamma,X} X^{[i]}_\gamma \|_F^2 = \sum_{J \in \mathcal{K}} \sum_{i=1}^{n_j} \frac{\sigma^{(J)}_i(X^{[i]}_\gamma)^2}{\sigma^{(J)}_1(X^{[i]}_\gamma)^2 + \gamma^{(i)} \gamma^{(i)}} \sum_{J \in \mathcal{K}} \text{rank}(X^{[J]}). \]

Though largely similar to the matrix case, there is however at least one difference as we discuss in Section 2. Due to its dependence on \( p = 0 \) and the family \( \mathcal{K} \), we abbreviate above algorithm (1.4) as IRLS-0\( \mathcal{K} \).

1.5. Data sparse optimization. With increasing dimensions \( d \), the size of the space \( \mathbb{R}^{n_1 \times \cdots \times n_d} \) quickly becomes prohibitively large. While for smaller instances, IRLS-0\( \mathcal{K} \) is by all means a viable algorithm, it otherwise remains a theoretical ideal. However, for hierarchical families \( \mathcal{K} \), that is if

\[ (J \subset S \lor S \subset J \lor J \cap S = \emptyset) \land J \neq S^c, \quad \forall J, S \in \mathcal{K}, \]

so called hierarchical decompositions [15] or, basically synonymously, tensor tree networks (cf. [10, 25]) provide remedy in the same way the ordinary low rank matrix decomposition does (cf. [26]). In the latter case, the data space \( \mathcal{D}_r := \{(Y,Z) \mid Y \in \mathbb{R}^{k_1 \times r}, \ Z \in \mathbb{R}^{r \times k_2} \} \) represents the low rank variety

\[ V_{\leq r}^{k_1,k_2} := \{ A \in \mathbb{R}^{k_1 \times k_2} \mid \text{rank}(A) \leq r \} \]

via the surjective (but not injective) bilinear map

\[ \tau_r : \mathcal{D}_r \to V_{\leq r}, \quad \tau_r(Y,Z) := YZ \in \mathbb{R}^{n_1 \times n_2}. \]

\[ 2 \] Though certainly interrelated, such are not matricizations of some common tensor.
The alternating method AIRLS then only requires to operate on \( D_r \), while directly minimizing \( f_\gamma \) subject to relaxed affine constraints (see Section 4). In the tensor case, where the rank becomes \( r = \{ \mathcal{r}(j) \}_{j \in \mathcal{K}} \in \mathbb{N}^\mathcal{K} \), the variety
\[
V^r_\mathcal{K} := \{ X \in \mathbb{R}^{n_1 \times \ldots \times n_d} \mid X[j] \in V^{n_r,j}_{\mathcal{r}(\mathcal{r}(j))}, J \in \mathcal{K} \},
\]
has a logarithmicly lower dimension and is likewise represented by a data space \( D_r \) together with a simple, surjective and multilinear contraction map \( \tau_r : D_r \to V^r_\mathcal{K} \) (see Section 5). Thereby, a sparse optimization as for matrices is also possible in higher dimensions. Ultimately, also AIRLS-0\( \mathcal{K} \) (see Section 6) distinguishes itself from well known unregularized alternating least squares (ALS) [20] only through an additional penalty term. However, it thereby not only becomes stable by means of [16], but it is derived from and directly minimizes the objective function \( f^\mathcal{K}_\gamma \) restricted to \( V^r_\mathcal{K} \).

2. Underlying structure and global behavior

Phrased more generalized, we in principle desire to solve the problem (cf. [26]) of finding
\[
X^* \in \arg\min_{X \in \mathcal{L}^{-1}(\gamma)} \mathcal{C}_\mathcal{V}(X), \quad \mathcal{C}_\mathcal{V}(v) := \min_{v \in \mathcal{V}} \dim(V),
\]
where in this setting the family of varieties \( \mathcal{V} \) is
\[
\mathcal{V}^r_\mathcal{K}_{\leq} := \{ V^r_\mathcal{K} \subset \mathbb{R}^{n_1 \times \ldots \times n_d} \mid r = \{ \mathcal{r}(j) \}_{j \in \mathcal{K}} \in \mathbb{N}^\mathcal{K} \},
\]
for \( V^r_\mathcal{K} \) as defined in (1.7). In general however, the dimension of \( V^r_\mathcal{K} \) does not equal \( \sum_{J \in \mathcal{K}} \mathcal{r}(J) \), and is thus not directly represented by the sum of ranks as in (1.1). While
\[
V^r_\mathcal{K}_{\leq} \subset V^r_\mathcal{K}_{\leq} \quad \Rightarrow \quad \mathcal{r}(J) \leq \mathcal{r}(J), \quad J \in \mathcal{K}, \quad \mathcal{r} \neq \mathcal{r} \quad \Rightarrow \quad \dim(V^r_\mathcal{K}_{\leq}) \leq \dim(V^r_\mathcal{K}_{\leq}),
\]
either of the converse implications holds true in general. Firstly, some differently indexed varieties are equal since some constellations \( r \in \mathbb{N}_0^\mathcal{K} \) are unfeasible [25].

Definition 2.1. The values \( r = \{ \mathcal{r}(j) \}_{j \in \mathcal{K}} \) are called (un)feasible (for \( n \in \mathbb{N}^d \)), if there exists (not) at least one tensor \( X \in \mathbb{R}^{n_1 \times \ldots \times n_d} \) with \( \text{rank}(X[j]) = \mathcal{r}(j) \), \( J \in \mathcal{K} \).

For hierarchical sets \( \mathcal{K} \), these bounds are (cf. [25]) \( r^{(\ell)}_{\mathcal{J}^\ell} \leq n_{\ell} \prod_{\ell \in \mathcal{E}_{\mathcal{K}}} r^{(\ell)}_{J^\ell} \) for \( \hat{\ell} \in \mathcal{E}_{\mathcal{K}}, v \in \mathcal{V} \). This natural interrelation of ranks is somewhat beneficial to the simplified sum-of-ranks approach as it excludes some extremal cases. The sum-of-ranks minimization is itself a necessary relaxation of the (arguably) more desirable objective function \( \mathcal{C}_\mathcal{V}^r \), yet it is closer than it might first seem. What remains however is that, contrarily to the matrix case, the varieties are only partially nested.

2.1. Determinant expansion and convergence of (global) minimizers.

Following from the matrix case, one can likewise expand the function \( f^r_\mathcal{K} \) into squared sums of minors defined as
\[
\det^2_k(A) := \sum_{I \in \mathcal{P}_k(n_{\ell})} \sum_{J \in \mathcal{P}_k(n_{\ell})} \det(A_{I,J})^2, \quad k = 1, \ldots, n_{J},
\]
for \( A_{I,J} := \{ A_{i,j} \}_{i \in I, j \in J} \in \mathbb{R}^{[I] \times [J]} \) and \( \mathcal{P}_k([\ell]) := \{ I \subseteq \{ 1, \ldots, \ell \} \mid |I| = k \} \). For simplicity of notation, we further define \( \det^2_k(A) := 1 \).

Corollary 2.2. Let \( X \in \mathbb{R}^{n_1 \times \ldots \times n_d} \) and \( \gamma \geq 0 \). Then
\[
\prod_{J \in \mathcal{K}} \prod_{i=1}^{n_{J}} (\sigma_i^{(J)}(X)^2 + \gamma) = \prod_{J \in \mathcal{K}} \sum_{k=0}^{n_{J}} \gamma^{(n_{J} - k)} \cdot \det^2_{k,J}(X[j]) = \sum_{s=0}^{\sum_{J \in \mathcal{K}} n_{J} - n} \gamma^s g_s(X)
\]
with
\[ g_s(X) := \sum_{(k^J) \in \Xi_s} \prod_{J \in \mathcal{K}} \det_{k^J}^2(X[^J]), \]  
for \( \Xi_s := \{ (k^J)_{J \in \mathcal{K}} \mid 0 \leq k^J \leq n_J, J \in \mathcal{K}, \sum_{J \in \mathcal{K}} k^J = s \} \).

**Proof.** As \( \prod_{i=1}^{n_J} (\sigma_i^{(J)}(X)^2 + \gamma) = \det(X[^J](X[^J])^T + \gamma I) \), the first equality follows by [26]. The third term is merely a restructured version. \( \square \)

The minimizers of these functions are nested in the sense of the following Lemma.

**Lemma 2.3.** For \( g_s(X), s = 0, \ldots, \sum_{J \in \mathcal{K}} n_J \), as in Corollary 2.2, we have
\[ g_s(X) = 0 \iff \sum_{J \in \mathcal{K}} \text{rank}(X[^J]) < s \]  
for all \( X \in \mathbb{R}^{n_1 \times \ldots \times n_d} \).

**Proof.** By definition of \( g_s(X) \), we have
\[ g_s(X) \neq 0 \iff \exists (k^J)_{J \in \mathcal{K}} : \sum_{J \in \mathcal{K}} k^J = s \ \forall J \in \mathcal{K} : \text{rank}(X[^J]) \geq k^J \]
\[ \iff \sum_{J \in \mathcal{K}} \text{rank}(X[^J]) \geq s. \]  \( \square \)

By Lemma 2.3, it directly follows that each \( g_s(X) = 0 \) implies \( g_{s+1}(X) = 0 \). With this structure, we can apply the nested minimization scheme as in [26] to conclude the following Theorem 2.4.

**Theorem 2.4.** Let
\[ s^* = \min_{X \in \mathcal{L}^{-1}(y)} \sum_{J \in \mathcal{K}} \text{rank}(X[^J]) \]  
Then for any convergent sequence of (global) minimizers \( X_\gamma \) of \( f_\gamma^K(X) \) subject to \( \mathcal{L}(X) = y \), we have
\[ X^* := \lim_{\gamma \to 0} X_\gamma \in \arg\min_{X \in \mathcal{L}^{-1}(y)} \sum_{J \in \mathcal{K}} \text{rank}(X[^J]) = s^* \prod_{J \in \mathcal{K}} \prod_{i=1}^{\text{rank}(X[^J])} \sigma_i^{(J)}(X) \]  
with
\[ \sigma_{\text{rank}(X[^J])+1}(X_\gamma)^2 \in \mathcal{O}(\gamma), \quad J \in \mathcal{K}. \]

If there is only one \( X_{s^*} \in \mathcal{L}^{-1}(y) \) with \( \sum_{J \in \mathcal{K}} \text{rank}(X_{s^*}[^J]) = s^* \), then \( X_\gamma \to X_{s^*} \).

**Proof.** Since \( \arg\min_{X \in \mathcal{L}^{-1}} g_s(X) \subset \arg\min_{X \in \mathcal{L}^{-1}} g_{s+1}(X) \) due to Lemma 2.3, the proof is analogous to the corresponding one in [26]. \( \square \)

3. Log-det tensor iteratively reweighted least squares (IRLS-0\( \mathcal{K} \))

Although the global minimizers of \( f_\gamma^K \) yield the sought solution, it is not practicable to directly minimize these functions or to find its extremal points. As in the matrix case, the map is augmented. While one here requires to introduce one weight for each \( J \in \mathcal{K} \), most results for the matrix case transfer directly due to the similar structure.
3.1. *Minimization of an augmented function.* The augmented map analogous to $f_\gamma$ corresponding to the tensor function $J^K_\gamma$ is

$$J^K_\gamma(X, \{W^{(J)}\}_{J \in \mathcal{K}}) := \sum_{J \in \mathcal{K}} J_{\gamma, m}(X^{[J]}, W^{(J)})$$

for

$$J_{\gamma, m}(A, H) = \text{trace}(H(ATA^T + \gamma I)) - \log \det(H) - m$$

$$= \sum_{J \in \mathcal{K}} \|H^{1/2}A\|_F^2 + \gamma \|H^{1/2}\|_F^2 - \log \det(H) - m,$$

where each $W^{(J)} \in \mathbb{R}^{n_J \times n_J}$ ranges over $W^{(J)} = (W^{(J)})^T \succ 0$ (symmetric positive definite). Consequently, with the same argumentation as in [11, 26], it is

$$\frac{\partial}{\partial W^{(J)}} J^K_\gamma(X, \{W^{(J)}\}_{J \in \mathcal{K}}) = X^{[J]}(X^{[J]})^T + \gamma I - (W^{(J)})^{-1}$$

and thus

$$W^{(J)}_{\gamma, X} := \arg\min_{W^{(J)} = (W^{(J)})^T \succ 0} J^K_\gamma(X, \{W^{(J)}\}_{J \in \mathcal{K}}) = (X^{[J]}(X^{[J]})^T + \gamma I)^{-1}.$$

It likewise holds true that

$$J^K_\gamma(X) = J^K_\gamma(X, \{W^{(J)}_{\gamma, X}\}_{J \in \mathcal{K}}).$$

Further, the minimizer in $X$ is determined by an ordinary least squares problem. In order to derive the closed form solution for the minimizer, we note that each $W^{(J)}, J \in \mathcal{K}$, defines linear operations

$$(W^{(J)})^\alpha : \mathbb{R}^{n_1 \times \ldots \times n_d} \rightarrow \mathbb{R}^{n_1 \times \ldots \times n_d}, \quad ((W^{(J)})^\alpha(X))^{[J]} := (W^{(J)})^\alpha X^{[J]}, \quad \alpha > 0.$$

We can thereby write

$$\sum_{J \in \mathcal{K}} \|(W^{(J)})^{1/2}X^{[J]}\|_F^2 = \sum_{J \in \mathcal{K}} \|(W^{(J)})^{1/2}(X)\|_F^2 = \|\tilde{W}^K(X)\|_F^2,$$

where $\tilde{W}^K(X) := \{(W^{(J)})^{1/2}(X)\}_{J \in \mathcal{K}} \in \mathbb{R}^{n_1 \times \ldots \times n_d \times |\mathcal{K}|}$. Based on the operator $\tilde{W}^K$ (cf. [26]), the sought minimizer is given by

$$X^K_W := \arg\min_{X \in \mathcal{L}^{-1}(y)} J^K_\gamma(X, \{W^{(J)}\}_{J \in \mathcal{K}}) = \tilde{W}^{-1} \circ \mathcal{L}^* \circ (\mathcal{L} \circ \tilde{W}^{-1} \circ \mathcal{L}^*)^{-1}(y)$$

for

$$\tilde{W}^K(X) := (\tilde{W}^K)^* \circ \tilde{W}^K(X) = \sum_{J \in \mathcal{K}} W^{(J)}(X),$$

where $(\cdot)^*$ denotes adjoint operators. Further, following [11, 26, 29], we have

$$\tilde{W}^K(X^K_W) \perp \text{kernel}({\mathcal{L}}).$$

Vice versa, $X^K_W$ is the unique solution to (3.5) subject to $\mathcal{L}(X^K_W) = y$. A more stable update formula is provided by [26] through

$$X^K_W = X_0 - \mathcal{K} \circ (\mathcal{K}^* \circ \tilde{W}^K \circ \mathcal{K})^{-1} \circ \mathcal{K}^* \circ \tilde{W}^K(X_0),$$

where $X_0$ is one arbitrary solution to $\mathcal{L}(X_0) = y$ and $\mathcal{K} : \mathbb{R}^{\prod_{\ell=1}^d n_\ell - t} \rightarrow \mathbb{R}^{n_1 \times \ldots \times n_d}$ is a kernel representation of $\mathcal{L}$, whereby image($\mathcal{K}$) = kernel($\mathcal{L}$). Due to the sum structure, also the gradient properties generalize to the tensor case.

---

3Due to the distinguishable roles of $J \in \mathcal{K}$ and the map $J^K_\gamma$, we here remain faithful to prior literature as for both the letter $J$ has been used before.
Corollary 3.1. It is
\[
\nabla_X f^K_{\gamma}(X) = \nabla_X J^K_{\gamma}(X, \{W^{(J)}\}_{J \in \mathcal{K}})|_{W^{(J)} = W^{(J)}_{\gamma,X}, \ J \in \mathcal{K}}.
\]
Thus \(X\) is a stationary point of \(f^K_{\gamma}\) if and only if \(X = X^K_{\gamma,\mathcal{W}}\) for \(W^{(J)} = W^{(J)}_{\gamma,X}, \ J \in \mathcal{K},\) which means that \((X, \{W^{(J)}_{\gamma,X}\}_{J \in \mathcal{K}})\) is a stationary point of \(J^K_{\gamma}\).

As in the matrix case, \(\gamma \to \infty\) provides a unique, canonical starting value.

Corollary 3.2. Independently of \(X^{(0)} \in \mathcal{L}^{-1}(y)\), it holds
\[
\lim_{\gamma \to \infty} \argmin_{X \in \mathcal{L}^{-1}(y)} f^K_{\gamma}(X) = \lim_{\gamma \to \infty} X^{(0)} \{W^{(J)}_{\gamma,X} \}_{J \in \mathcal{K}} = \argmin_{X \in \mathcal{L}^{-1}(y)} \|X\|_F,
\]
where the first limit is possibly a set convergence.

3.2. Complementary weights. In the matrix case [26], there is one more equitable choice \(f^{(2)}(A) = \log \det(A^T A + \gamma I)\) as opposed to \(f^{(1)}(A) = f_{\gamma}(A) = \log \det(A A^T + \gamma I)\). For families \(\mathcal{K}\) containing more subsets, each set \(J \in \mathcal{K}\) may be replaced by its complement. For a subset \(S \subset \mathcal{K}\), let therfore
\[
\mathcal{K}^{S} := (\mathcal{K} \setminus S) \cup \{J^c \ | \ J \in S\}, \quad J^c := [d] \setminus J,
\]
for \(W^{(J)} = W^{(J)}_{\gamma}, \ J \in \mathcal{K}^{S}\). Although the updates \(X^{K}_{W\gamma X}\) and \(X^{K^{S}}_{W\gamma X}\) in general differ, the overall properties outlined in Section 3 are not influenced as
\[
f^K_{\gamma}(X) = \sum_{J \in S} \sum_{i=1}^{n_J} \log(\sigma^2_i)(X)^2 + \gamma + \sum_{J \in \mathcal{K} \setminus S} \sum_{i=1}^{n_J} \log(\sigma^2_i)(X)^2 + \gamma
\]
\[
= f^K_{\gamma}(X) + \sum_{J \in S} (n_J - n_J) \log \gamma.
\]
While the weights are in that sense interchangeable, switching between complementary weights becomes essential for AIRLS-0\(\mathcal{K}\) as captured in Lemma 6.2.

3.3. Adjusted IRLS-0\(\mathcal{K}\) algorithm. Based on a monotonically declining sequence \(\{\gamma^{(i)}\}_{i \geq 0} \subset \mathbb{R}_{>0}\) (cf. Definition 1.1), and (optionally) a sequence \(S_i \subset \mathcal{K}\) (cf. Section 3.2), Algorithm 1 defines the sequence \(\{(X^{(i)}, \{W^{(i,J)}\}_{J \in \mathcal{K}})\}_{i \geq 0}\) with \(\mathcal{L}(X^{(i)}) = y\) and \((W^{(i,J)})^T = W^{(i,J)} > 0, \ J \in \mathcal{K}, \ i \geq 0\). These iterates behave largely analogously to the matrix version [26] (cf. [5,7,11,29]). In particular, that case is included in Theorem 3.3 for \(d = 2\) and \(\mathcal{K} = \{\{1\}\}\).

Algorithm 1 Iteratively reweighted least squares with switching weights

1: set \(X^{(0)} \in L^{-1}(y), \ \gamma^{(0)} > 0\)
2: for \(i = 1, 2, \ldots\) do
3: set \(S_{i-1} \subset \mathcal{K}\) (cf. Section 3.2)
4: \(\{W^{(i-1,J)}\}_{J \in \mathcal{K} \setminus S_{i-1}} := \{W^{(i,J)}_{\gamma^{(i-1)}, X^{(i-1)}}\}_{J \in \mathcal{K} \setminus S_{i-1}}\) (cf. (3.1))
5: \(X^{(i)} := X^{K_{S_{i-1}}}_{W^{(i-1)}}\) (cf. (3.3))
6: set \(\gamma^{(i)} \leq \gamma^{(i-1)}\)
7: end for

Theorem 3.3. Let \(\{(X^{(i)})\}_{i \geq 0}\) be generated by Algorithm 1 for \(\{S_i\}_{i \in \mathbb{N}_0}\) and the weakly decreasing sequence \(\{\gamma_i\}_{i \geq 0} \subset \mathbb{R}_{>0}\). Let further \(\mathcal{S}_i^* \subset \mathcal{L}^{-1}(y)\) be the stationary points of \(f^K_{\gamma}(X^{(i)})\) for \(\gamma > 0\), as well as \(\mathcal{S}_\infty := \lim_{i \to \infty} \mathcal{S}_i^*\).

(i) For each \(i \in \mathbb{N}\) and each \(S \subset \mathcal{K}\), it holds
\[
f^K_{\gamma^{(i)}}(X^{(i)}) \leq f^K_{\gamma^{(i-1)}}(X^{(i-1)}).
\]
(ii) If $\gamma^* > 0$, then the sequences $X^{(i)}$ and $|f^{\gamma^*}_{\gamma^*} (X^{(i)})|$, $\mathcal{S} \subset \mathcal{K}$, remain bounded.

(iii) Further, if $\gamma^* > 0$, then

\[
\lim_{i \to \infty} \|X^{(i)} - X^{(i-1)}\|_F = 0
\]

and each accumulation point of $X^{(i)}$ is in $\mathcal{S}^*_{\gamma^*}$.

(iv) (See Remark 3.4) Let $\Theta \subset \mathbb{R}_{>0}$ be an arbitrary, infinite, bounded set with its only accumulation point at $\inf(\Theta) = 0$, and let $\delta_i := \inf_{S \in \mathcal{S}^*_{\gamma(i)}} \|X^{(i)} - S\|$, $i \in \mathbb{N}$.

For an arbitrary, bounded sequence $A = \{a_i\}_{i \in \mathbb{N}_0}$ with $\inf(A) > 0$ (e.g. $a_i = 1$, $i \in \mathbb{N}_0$) and for $\gamma^{(0)} = \max(\Theta)$, we recursively define

\[
\gamma^{(i+1)} = \begin{cases} 
\theta_i & \text{if } \alpha_i \delta_i < \theta_i, \\
\gamma^{(i)} & \text{otherwise}
\end{cases}
\]

\[
\theta_i := \max \{ z \in \Theta \mid z < \gamma^{(i)} \}, \quad i \in \mathbb{N}_0.
\]

Then $\lim_{i \to \infty} \delta_i = \gamma^* = 0$ and for at least one subsequence $\{X^{(i)}\}_{i \in \mathbb{N}}$, there exists a sequence of stationary points $\{S_i\}_{i \in \mathbb{N}}$, $S_i \in \mathcal{S}^*_{\gamma^{(i)}}$, with $\|S_i - X^{(i)}\| \to 0$.

Remark 3.4. Part (iv) of Theorem 3.3 as well as its proof are literally the same as in the matrix case [26]. Roughly, if the sequence $\{\gamma^{(i)}\}_{i \in \mathbb{N}}$ is decreased to $\gamma^* = 0$ slowly enough, then $X^{(i)}$ can only converge to a limit of stationary points of $f_\gamma|_{\mathcal{L}^{-1}(y)}$ for $\gamma \searrow 0$. The contrary case of too fast decline has been covered in [26] as well.

Proof. See Appendix A. \qed

4. RELAXED ITERATIVELY Reweighted LEAST SQUARES

Too large mode sizes $n$ or high dimensions $d$ in practice prohibit to even operate on the spaces $\mathcal{L}^{-1}(y)$ or $\mathbb{R}^{n_1 \times \cdots \times n_d}$ directly. As hinted on in Section 1.5, so called hierarchical decompositions can provide remedy in the same way low rank matrix decompositions do. This however first requires to relax the affine constraint $\mathcal{L}(X) = y$.

4.1. Relaxation of affine constraint. Let $a_\gamma(s) := s - \sum_{J \in \mathcal{K}} n_J \log(\gamma)$, $\gamma > 0$. As each of these function is monotonically increasing, a composition with such does not change minimizers. We correspondingly define

\[
(f^{a_\gamma}_{\gamma})_{\gamma}^{\mathcal{K}}(X) := a_\gamma \circ \mathcal{L}^{\gamma} = \log \prod_{J \in \mathcal{K}} \prod_{i=1}^{\infty} \left( 1 + \frac{\sigma^{(J)}_i(X)}{\gamma} \right),
\]

with $\sigma^{(J)}_i(X) := 0$ for $i > n_J$, $J \in \mathcal{K}$. Likewise, let $J^{a_\gamma}_{\gamma} \mathcal{K}(X, \{W^{(J)}\}_{J \in \mathcal{K}}) := a_\gamma \circ J^{\gamma}_{\gamma} \mathcal{K}(X, \{W^{(J)}\}_{J \in \mathcal{K}})$. With the same reasoning as in [26], one then defines

\[
E^{a_\gamma}_{\gamma} \mathcal{K}(X) := \| \mathcal{L}(X) - y\|_F^2 + c_\gamma \cdot \omega^2 \cdot (f^{a_\gamma}_{\gamma})_{\gamma}^{\mathcal{K}}(X),
\]

\[
J^{a_\gamma}_{\gamma} \mathcal{K}(X, W) := \| \mathcal{L}(X) - y\|_F^2 + c_\gamma \cdot \omega^2 \cdot J^{a_\gamma}_{\gamma} \mathcal{K}(X, W),
\]

for an appropriate scaling constant $c_\gamma$. As $\frac{\partial}{\partial \gamma} (f^{a_\gamma}_{\gamma} \mathcal{K}(X) = c_\gamma \cdot \frac{\partial}{\partial \gamma} f^{a_\gamma}_{\gamma} \mathcal{K}(X)) \geq 0$, the choice $\omega = \sqrt{\gamma}$ seems suitable. In that case, we skip the index $\omega$. 

ASYMPTOTIC LOG-DET SUM-OF-RANKS MINIMIZATION VIA TENSOR (A)IRLS 9
Algorithm 2 Subspace restricted IRLS with switching weights

1: set $X^{(0)} \in \mathcal{L}^{-1}(y)$, $\gamma^{(0)} > 0$
2: for $i = 1, 2, \ldots$ do
3: set $S_{i-1} \subseteq \mathcal{K}$ (cf. Section 3.2)
4: $\{W(i-1,J)\}_{J \in \mathcal{K}(S_{i-1})} := \{W_{\gamma^{(i-1)}}^{(J)}\}_{J \in \mathcal{K}(S_{i-1})}$ (cf. (3.1))
5: set a subspace $T_{i-1} \subseteq \mathbb{R}^{n_1 \times \ldots \times n_d}$ with $T_{i-1} \ni X^{(i-1)}$
6: $X^{(i)} := \arg\min_{X \in T_{i-1}} J_{\gamma^{(i-1)}}(X, \{W^{(i-1,J)}\}_{J \in \mathcal{K}(S_{i-1})})$ (cf. (4.2))
7: set $\gamma^{(i)} \leq \gamma^{(i-1)}$
8: end for

4.2. Subspace dependent, relaxed optimization algorithm. To later incorporate the alternating optimization, we here also consider an additional sequence of subspaces $\{T_i\}_{i \in \mathbb{N}}$, with $T_i \subseteq \mathbb{R}^{n_1 \times \ldots \times n_d}$, as well as $T_i \cap T_{i-1} \ni X^{(i)}$, $i \in \mathbb{N}_0$. The latter condition ensures that the previous iterate remains admissible. This then yields the modified Algorithm 2. While the objective function is still monotonically decreased as provided by Corollary 4.1, to show the remaining parts of Theorem 3.3 as far as possible for now remains subject to future research.

Corollary 4.1. For $X^{(i)}$ as defined by Algorithm 2 it holds

$$F_{\gamma^{(i)};\mathcal{K}}(X^{(i)}) \leq F_{\gamma^{(i-1)};\mathcal{K}}(X^{(i-1)}),$$

for all $i \in \mathbb{N}$ and all $\mathcal{S} \subset \mathcal{K}$.

Proof. The argumentation is the same as in Theorem 3.3 part (i) as steps (a) to (g) analogously hold true (cf. Section 4.1).

5. Hierarchical decomposition

We briefly reintroduce hierarchical tensor decompositions [15] as tensor tree networks with reference to the introductory Section 1.5. For further reading, we recommend [8, 10, 15, 17, 23, 25, 30].

5.1. Notational deviation. In the following, $G = (V, E)$ denotes a tree graph with vertices $V \supseteq \{d\}$ and edges $E \subseteq \{(v, w) : v \neq w \in V\}$. Due to the complex description of general tensor (tree) networks, we require a certain minimum of notational deviation. That is, we dismiss the order of modes when indexing tensors. Instead, in order to avoid ambiguity, each specific object is consistently referenced with the same, distinctly assigned labels, based on the graph $G = (V, E)$. The first group is given by $\alpha_S = \{\alpha_\mu\}_{\mu \in S}$, for $\alpha_S \in [n_S]$, $n_S = \prod_{\mu \in S} n_\mu$, $S \subseteq V$. We set $n_\mu = 1$ for $\mu > d$, but any such $\alpha_\mu$ is only denoted when required for notational simplicity. Further, the second group is given by $\beta = \{\beta_e\}_{e \in E}$ with $\beta_e \in [r^{(L_e)}]$, $J_e \in \mathcal{K}$ (see Section 5.3), whereas the measurement index is denoted by $\zeta \in [\ell]$. For each such label, we correspondingly define the spaces

$$\delta_{\alpha_v} := \mathbb{R}^{[n_v]}, \quad \delta_{\beta_e} := \mathbb{R}^{[r^{(L_e)}]}, \quad \delta_{\zeta} := \mathbb{R}^{[\ell]}.$$

The entirety of labels is formally required to be ordered, but the exact ordering is irrelevant. To each collection $\Gamma$ of such labels, we consequently assign the space

$$\delta_{\Gamma} := \bigotimes_{\gamma \in \Gamma} \delta_{\gamma}.$$  

(5.1)

Some, in particular labels corresponding to edges also appear as unequally treated, so called primed labels $\beta'_e \neq \beta_e, e \in E$. Each is however still thought to refer to the same, implicitly declared positions of its unprimed twin. Throughout this section,
it shall become apparent that it is in fact mostly redundant to explicitly denote these labels. While we nevertheless here hold on to indices, Section SM5 does make use of this fact to more compactly repeat some of following statements and lay out their proofs. What is here introduced as notation, is the basis to the formalized arithmetic introduced in [25]. For the MATLAB toolbox that realizes the latter through automated contractions, on which the implementation of (A)IRLS-0K is based on, please contact the author.

5.2. Graph notation. We denote each the path from excluding \( c \in V \) to excluding \( v \in V \setminus \{v\} \) within a tree \( G = (V,E) \) as the unique ordered set
\[
(5.2) \quad c \mapsto v := (p_1, \ldots, p_{-1}) = p \subset V,
\]
for which \( \{c, p_1\} \in E, \{p_i, p_{i+1}\} \in E, i = 1, \ldots, |p| - 1 \) as well as \( \{p_{-1}, v\} \in E \).

We further define the neighbors of \( v \in V \), as well as the predecessor and set of descendants of \( v \in V \setminus \{c\} \) relative to \( c \in V \) as

\[ \text{neigh}(v) := \{ b \in V \mid \{b, v\} \in E \}; \quad \text{pred}_c(v) := p_{-1}, \quad \text{desc}_c(v) := \text{neigh}(v) \setminus \{p_{-1}\}. \]

We define the branches relative to \( c \in V \) as
\[ \text{branch}_c(v) := \{v\} \cup \{b \in V \setminus \{c, v\} \mid v \in c \mapsto b\}. \]

Each root \( c \in V \) splits the graph into the multiple connected components of \( V \setminus \{c\} \),
\[
\bigcup_{h \in \text{neigh}(c)} \text{branch}_c(h) = V \setminus \{c\}.
\]

For any \( v \neq w \in V \), we further define the sets \( J_w(v) := \text{branch}_w(v) \cap [d] \). Thus if \( e = \{v, w\} \in E \) is an edge, then \( J_w(v) \cup J_v(w) = [d] \).

5.3. Tree corresponding to hierarchical family. Without loss of generality, we from here on postulate that hierarchical families \( \mathcal{K} \) (cf. Section 1.5) are by definition also dimension separating. That is, we assume that there does not exist a map \( \pi : [d] \to [d - 1] \), for which \( \pi(J) \notin \{\pi(J), [d - 1] \setminus \pi(J)\} \) for all \( J \subseteq \mathcal{K} \).

**Lemma 5.1.** Each (dimension separating) hierarchical family \( \mathcal{K} \) defines an, up to equivalence, unique tree \( G_\mathcal{K} = (V,E), \ V \supseteq [d] \) and root \( c \in V \), for which \( |E| = |V| - 1 = |\mathcal{K}| \) and \( \mathcal{K} = \{J_e(v)\}_{v \in V \setminus \{c\}} \) — and vice versa.

**Proof.** See for example [15, 25].

**Definition 5.2.** Let \( G_\mathcal{K} \) correspond to the hierarchical family \( \mathcal{K} \). We define \( J_e \in \{J_w(v), J_v(w)\} \), \( e = \{v, w\} \in E \), as each the one set that is contained in \( \mathcal{K} \).

This convention implies a bijection \( \mathcal{K} = \{J_e \mid e \in E\} \) to \( E \). The simple graph that corresponds to the matrix case \( \mathcal{K}_2 = \{\{1\}\} \) for \( d = 2 \) is for instance given by the tree
\[
G_{\mathcal{K}_2} = (V,E), \quad V = \{1, 2\}, \quad E = \{\{1, 2\}\},
\]
whereby \( J_{\{1, 2\}} = \{1\} \). For \( \mathcal{K}_{\text{Tucker}} = \{\{1\}, \ldots, \{d\}\} \) (cf. Example 5.5), we have
\[
(5.3) \quad G_{\mathcal{K}_{\text{Tucker}}} = (V,E), \quad V = \{1, \ldots, d + 1\}, \quad E = \{\{1, d + 1\}, \ldots, \{d, d + 1\}\},
\]
and \( J_{\{\mu, \mu + 1\}} = \{\mu\}, \mu \in [d] \). As required later, for subsets \( S \subseteq V \), we further define
\[
(5.4) \quad E_S := \{\{v, w\} \in E \mid v \in S, \ w \in \text{neigh}(v)\}, \quad \tilde{E}_S := \{\{v, w\} \in E \mid v, w \in S\}, \ \partial E_S := E_S \setminus \tilde{E}_S.
\]
For \( S = \{v\}, \ v \in V \), we may skip set brackets. Thus, \( \tilde{E}_v = \{\{v, h\}\}_{h \in \text{neigh}(v)} \).
5.4. Representation map corresponding to tree. Whereas each hierarchical family $\mathcal{K}$ defines a tree $G_{\mathcal{K}} = (V, E)$, each such (not necessarily rooted) graph together with $r \in \mathbb{N}^\mathcal{K}$ in turn defines a certain data space $D_r$ and a representation map $\tau_r : D_r \to \mathbb{R}^{n_1 \times \cdots \times n_d}$ for values $r \in \mathbb{N}^\mathcal{K}$.

**Definition 5.3.** With reference to Section 5.1, let

$$D_r := \bigtimes_{v \in V} \mathcal{S}_{m_v}, \quad m_v := \{\beta^v\}_{\beta \in E_v} \cup \{\alpha_v\} \quad \text{if } v \in [d],$$

$$\emptyset \quad \text{otherwise.}$$

The dimension of each node $N_v \in \mathcal{S}_{m_v}$, $\{N_v\}_{v \in V} \in D_r$, is thus the degree of $v \in V$, plus one if $v \in [d]$. The representation map $\tau_r$ is now defined as the map that proceeds each a contraction over modes with common labels. With the notation declared in Section 5.1, we may write

$$\tau_r(N)_{\alpha_1, \ldots, \alpha_d} := \sum_{\beta \in E \mu \in [d]} \prod_{v \in \mathcal{V} \setminus [d]} (N_v)_{\alpha_v, (\beta^v)_{e \in E_v}} \prod_{v \in [d]} (N_v)_{\beta^v_{e \in E_v}},$$

where $\alpha_{\mu} \in [n_{\mu}]$, $\mu \in [d]$.

**Example 5.4.** In the matrix case with $r = r^{(J(1, 2))} \in \mathbb{N}$, we simply have an ordinary matrix multiplication (cf. Section 1.5) $\tau_r(Y, Z)_{\alpha_1, \alpha_2} = \sum_{\beta_1=1}^{\beta_2} Y_{\alpha_1, \beta_1} Z_{\beta_1, \alpha_2}$, where the summation ranges over $\alpha_1 \in [n_1]$ and $\alpha_2 \in [n_2]$. Here, $\beta = \beta^{(1, 2)} \in [r]$ is the label assigned to the only edge.

**Example 5.5.** For $d \in \mathbb{N}$, the Tucker format [40] or MLSVD [9] is defined through the graph $\mathcal{K}_{\text{Tucker}}$ (5.3) and consists of the components $\{N_{\mu}\}_{\mu=1}^{d+1} \in D_r$ of sizes $N_{\mu} \in \mathbb{R}^{n_{\mu} \times \ldots \times n_{d+1}}$ and $N_{d+1} \in \mathbb{R}^{\times \ldots \times \times n_{d+1}}$. The corresponding contraction map is given by (though less convenient when written out in particular cases)

$$X_{\alpha_1, \ldots, \alpha_d} = \tau_r(N_1, \ldots, N_d, N_{d+1})_{\alpha_1, \ldots, \alpha_d}$$

$$= \sum_{\beta^{(1, d+1)}=1}^{\beta^{(d, d+1)}} \sum_{\beta^{(d, d+1)}=1}^{\beta^{(d, d+1)}} (N_1)_{\alpha_1, \beta_{(1, d+1)}} \ldots (N_d)_{\alpha_d, \beta_{(d, d+1)}} (N_{d+1})_{\beta_{(1, d+1)}, \ldots, \beta_{(d, d+1)}},$$

for $\alpha_{\mu} = 1, \ldots, n_{\mu}$, $\mu = 1, \ldots, d$ as visualized in Fig. 1.

![Figure 1](https://via.placeholder.com/150)

**Figure 1.** [Left] The contraction diagram for the Tucker representation in Example 5.5 for $d = 4$. The dotted line indicates the part which for $J = \{1\}$ yields $Z^{(1)}$, whereas $Y^{(J)} = N_1$ (cf. (5.7)). [Right] A balanced binary hierarchical Tucker (HT) representation (cf. Section 5.6) for the exhaustive family $\mathcal{K} = \{1, 2\}, \{1\}, \ldots, \{4\}$, $Y^{((1, 2))} = \tau_r((N_1, N_2, \tilde{N}_{d+1}))$ (cf. (5.6)). In contrast to conventional literature (cf. [15]), the root node has been omitted as it is redundant here (cf. [25]).

---

\[\text{subject to further orthonormality constraints (cf. Section 5.7})\]
While initially defined on the whole network, we can also extend the map \( \tau_r \) to contract nodes over any subset \( S \subset V \) via
\[
(5.6) \quad \tau_r(\{N_s\}_{s \in S})_{(\alpha_s)_{s \in S},(\beta^r)_{s \in \partial E_S}} := \sum_{\beta^r, \in \partial E_S} \prod_{v \in S} (N_v)_{\alpha_v, (\beta^r)_{s \in E_v}},
\]
for \( \alpha_s \in [n_s], \ s \in S, \) and \( \partial E_S \) as well as \( \hat{E}_S \) as defined by (5.4). Here, some \( \alpha_v, \) that is for \( v > d, \) are redundant (cf. Section 5.1).

5.5. **Decomposition theorem.** The following theorem is fundamental to hierarchical tensor approximation theory.

**Theorem 5.6** ([15]). Let \( G_K \) be the tree corresponding to the hierarchical family \( K \) (cf. Lemma 5.5). Then for each \( r = \{r(J)\}_{J \in K} \in \mathbb{N}^K \), the according multilinear representation map \( \tau_r : D_r \to \mathbb{R}^{n_1 \times \cdots \times n_d} \) is non-injective with image(\( \tau_r \)) = \( V_{\leq r}^K \).

In other words, for each tensor \( X \in \mathbb{R}^{n_1 \times \cdots \times n_d} \) with rank(\( X_{(J)} \)) \( \leq r(J), \ J \in K \), there exists a (non-unique) decomposition \( N \in D_r \) with \( X = \tau_r(N) \). Each edge \( e = [v,w] \in E \), assuming \( J = J_e = J_v(w) \in K \), splits the tree into two disconnected subgraphs and yields a corresponding matrix decomposition
\[
(5.7) \quad X_{(J)} = Y^{(J)}_{v} Z^{(J)}_{w}, \quad Y^{(J)}_{v} \in \mathbb{R}^{[n_j] \times r(J)}, \quad Z^{(J)}_{w} \in \mathbb{R}^{r(J) \times [n_r]}.
\]
The matrices \( Y^{(J)}_{v} \) and \( Z^{(J)}_{w} \) are obtained by contractions over each \( (N_h)_{h \in \text{branch}_w(v)} \) and \( (N_h)_{h \in \text{branch}_v(w)} \), respectively. In explicit, abbreviating \( S = \text{branch}_w(v) \), we have
\[
Y^{(J)}_{v,\alpha,\beta} = \tau_r(\{N_s\}_{s \in S})_{(\alpha_s)_{s \in J},(\beta^r)_{s \in E_v}} := \sum_{\beta^r, \in \partial E_V} \prod_{v \in S} (N_v)_{\alpha_v, (\beta^r)_{s \in E_v}},
\]
for \( \hat{E}_S \) as defined in Section 5.3. Given (5.7), it is easy to see that indeed image(\( \tau_r \)) \( \subseteq V_{\leq r}^K \), whereas the other direction requires some more work (cf. [15,25]).

**Lemma 5.7.** The dimension of the variety corresponding to a feasible \( r \in \mathbb{N}^K \) for a hierarchical family \( K \) is
\[
\dim(V_{\leq r}^K) = (\sum_{\mu \in [d]} n_{\mu} \prod_{e \in E_{\mu}} r(e)) + (\sum_{v \in V \setminus [d]} \prod_{e \in E_v} r(e)) - \sum_{e \in E} (r(e))^2,
\]
where \( G_K = (V,E) \) is the corresponding graph. The set \( V_{\leq r}^K \) in turn is a manifold of equal dimension.

**Proof.** Follows by a generalization of the argumentation in [21,41].

5.6. **Exhaustive hierarchical families.** The larger the family \( K \), the more regularizing the IRLS approach. Thus, one may desire such to be exhaustive in the following sense.

**Definition 5.8.** Let \( K \) be a hierarchical family. We say \( K \) is exhaustive if there does not exist another hierarchical family \( \tilde{K} \) with \( \tilde{K} \supseteq K \).

Exhaustive hierarchical families in a certain sense yield particularly data sparse formats as specified in the following Lemma 5.9. For any such family, it further holds \( |K| = 2d - 3 = |E| \) and \( |V| = 2d - 2 \) (cf. Lemma 5.1).

**Lemma 5.9.** Let \( K \) be an exhaustive hierarchical family. Then \( G_K \) consists only of inner vertices \( v \in V \setminus [d] \) of degree 3 and leaves \( v \in [d] \) of degree 1.

**Proof.** See for instance [15,25].

---

5The rank considered therein is implicitly assumed to be feasible.
The Tucker family $\mathcal{K}_{\text{Tucker}}$ for example is not exhaustive (for $d \geq 4$). The degree of the vertex $d + 1 \in V$ is $d$, whereby the dimension of the node $N_d+1$ is $d$ as well. For $d = 4$, all exhaustive families are equivalent (up to permutation of modes) to $\mathcal{K} = \{\{1,2\}, \{1\}, \{2\}, \{3\}, \{4\}\}$ (see Fig. 1). In general, exhaustive hierarchical families correspond to so called binary hierarchical Tucker formats (cf. [15, 25]).

5.7. Rooted trees and orthonormalization. A root $c \in V$, if at all, may be chosen freely, leading us back to the choice of complementary weights in Section 3.2.

**Lemma 5.10.** For each $c \in V$, there exists a unique subset $S_c \subset K$ for which $\mathcal{K}^{S_c} = \{J_c(v) \mid v \in V \setminus \{c\}\}$ (cf. Section 5.2).

*Proof.* Follows directly with $S_c = \{J_c(v) \mid J_c(v) \notin K, v \in V \setminus \{c\}\} \subseteq K$. □

The set equality in Lemma 5.10 implies that for each $J \in \mathcal{K}^{S_c}$, there is a unique vertex $v = \tau_{c,J} \in V \setminus \{c\}$ with $J = J_c(v)$. Note that only the sets $S_c$, $c \in V$, again lead to hierarchical families $\mathcal{K}^{S_c}$ as opposed to the $2^{\vert K\vert}$ generally possible subsets $S \subset K$. One can utilize the non-injectivity of the map $\tau_r$ to orthonormalize the representation in the sense of the following Theorem 5.11, yet without the need to calculate the represented, full tensor.

**Theorem 5.11.** Let $r(J) = \text{rank}(X^{[J]})$, $J \in K$. Then there exists a representation $X = \tau_r(N)$, $N \in D_r$, such that $Y(J) \in \mathbb{R}^{n \times J \times (J)}$, $J \in \mathcal{K}^{S_c}$, as defined in (5.7), are orthonormal matrices.

*Proof.* Can for example be found in [25]. □

Note that the matrices $Y(J)$ in Theorem 5.11 are defined via the representation $N$. If may further be achieved that these matrices each consist of the left singular vectors, $Y(J) = U(J)$, of the compact matrix SVDs $X^{[J]} = U(J)S(J)(V(J))^T$, $J \in \mathcal{K}^{S_c}$. Thereby, the decomposition in fact becomes essentially unique$^6$ [15, 25]. However, mere orthonormality is in general sufficient and can be ensured with significantly less effort in an alternating optimization. In case of the Tucker format (Example 5.5), if indeed $Y(J) = U(J)$, this canonical form is specifically known as MLSVD [9], while for the tensor train format [32], it is known as canonical MPS [42]. General canonical forms of tensor tree networks and their properties are further discussed in [25].

6. Alternating iteratively reweighted least squares (AIRLS-0\(K\))

In this section, let $K$ be a hierarchical family, $G_K = (V, E)$ the corresponding tree as well as $\tau_r : D_r \rightarrow V_k^X$, with $D_r = X_{v \in V} S_{m_v}$, the representation map for $r \in \mathbb{N}^K$ as described in Section 5. The idea of alternating least squares (ALS) is to in each step fixate $N = \{N_v\}_{v \in V} \in D_r$ but the one component $N_c$, where the root $c \in V$ iteratively cycles through all vertices. We therofore define the linear map

$$N_{\neq c} : S_{m_c} \rightarrow V^{X}_{\neq c}, \quad N_{\neq c}(\hat{N}_c) = \tau_r(\{N_v\}_{v \in V \setminus \{c\}} \cup \{\hat{N}_c\}).$$

As the image of that map is independent of the specific, chosen representation, we obtain the (well defined) subspace (cf. Section 4.2)

$$T_c(\tau_r(N)) := \text{image}(N_{\neq c}).$$

Though one avoids to ever calculate the full tensor $X = \tau_r(\{N_v\}_{v \in V}) \in \mathbb{R}^{n_1 \times \ldots \times n_d}$, we define the resulting update as

$$X_{\gamma,\omega}^{N_c} := \arg\min_{X \in T_c(N)} J^{\mathcal{K}^{S_c}}(X, \{W^{[J]}\}_{J \in \mathcal{K}^{S_c}}),$$

$^6$Essentially here refers to the same weak uniqueness as for the conventional matrix SVD.
as well as \((N_c)_{\gamma,\omega,W}^N, W \in \mathcal{S}_m\) via \(N_{\#,c}((N_c)_{\gamma,\omega,W}^N) := X_{\gamma,\omega,W}^N\). The subset \(\mathcal{S}_c \subseteq \mathcal{K}\) is defined as by Lemma 5.10, the objective functions in (4.2). The subsequent sections are summarized in Algorithm 3, though it generates the same iterates \(X^{(i)} = \tau_r(N^{(i)}), \ i \in \mathbb{N}_0\), as Algorithm 2 when the subspaces are chosen according to (6.1).

6.1. Sweeps, micro steps and stability. The update \(X_{\gamma,\omega,W}^N\) (cf. (6.2)) is independent of the specific, chosen representation \(N\) of the previous iterate \(X\) (cf. Theorem 5.6). Thus, for each \(c \in V\), the updating maps

\[
\mathcal{M}_r^{(c)} : \mathcal{D}_r \to \mathcal{D}_r, \quad \mathcal{M}_r^{(c)}(N) := \{N_c\}_{c \in V \setminus \{c\}} \cup \{(N_c)_{\gamma,\omega,W}^N\}, \quad W = W_{\gamma,\tau_r(N)},
\]

operating on the data space, as well as the one operating on the full tensor space,

\[
\zeta_{\mathcal{M}^{(c)}} : \mathbb{R}^{n_1 \times \ldots \times n_d} \to \mathbb{R}^{n_1 \times \ldots \times n_d}, \quad \zeta_{\mathcal{M}^{(c)}}(X) := \tau_r(X) \circ \mathcal{M}_r^{(c)}(X) \circ \tau_r^{-1}(X)(X),
\]

are well defined. Here, \(r(X) \in \mathbb{N}^N\) denotes the ranks of each \(X\) and \(N = \tau_r^{-1}(X)\) is each an arbitrary representation. A whole sweep (for fixed \(\gamma\) and \(\omega\)) is defined as

\[
\mathcal{M}_r := \bigcirc_{c \in V} \mathcal{M}_r^{(c)}, \quad \zeta_{\mathcal{M}} := \bigcirc_{c \in V} \zeta_{\mathcal{M}^{(c)}},
\]

where the order of composition may be chosen as most suitable. Issues around these functions in particular concerning stable rank adaptivity have been discussed in [16].

6.2. Representation based evaluation. In order to obtain a practically viable algorithm, it remains to show that each next iterate \(X_{\gamma,\omega,W}^N\) (cf. (6.2)) given \(W = W_{\gamma,\tau_r(N)}\) and \(\omega = \sqrt{\gamma} > 0\) (cf. Section 4.1) can indeed be calculated through its representation, that is, without the need to construct full tensors in \(\mathbb{R}^{n_1 \times \ldots \times n_d}\). The updated node \((N_c)_{\gamma,\omega,W}^N \in \mathcal{S}_m\), for which \(X_{\gamma,\omega,W}^N = N_{\#,c}((N_c)_{\gamma,\omega,W}^N)\), is given by the linear least squares problem (cf. Sections 3.1, 3.2 and 4.1)

\[
(N_c)_{\gamma,\omega,W}^N = \arg \min_{\tilde{N}_c \in \mathcal{S}_m} \|L \circ N_{\#,c}(\tilde{N}_c) - y\|^2 + \ell c \gamma \sum_{J \in \mathcal{K}_{\#,c}} \|W(J)^{1/2} N_{\#,c}(\tilde{N}_c)\|^2_F,
\]

for \(W(J)\) as in (3.4). The minimizer is thus given as solution \((N_c)_{\gamma,\omega,W}^N := \tilde{N}_c\) to

(6.3) \(N_{\#,c} \circ L^* \circ L \circ N_{\#,c}(\tilde{N}_c) + \ell c \gamma \sum_{J \in \mathcal{K}_{\#,c}} N_{\#,c} \circ W(J) \circ N_{\#,c}(\tilde{N}_c) = N_{\#} \circ L^*(y)\).

Following are two aspects that are required for a representation based evaluation. The first one in Section 6.3 depends on the operator \(L\) itself and can in that sense not be influenced. The second one in Section 6.4 in turn merely asks for the right choices of \(S_c\), namely the one in Lemma 5.10, and can thus always be achieved.

6.3. Decomposition of measurement operator. Like each linear operator, \(L : \mathbb{R}^{n_1 \times \ldots \times n_d} \to \mathbb{R}^\ell\) has a tensor description \(L \in \mathbb{R}^{\ell \times n_1 \times \ldots \times n_d}\) in terms of

(6.4) \(L(X)_\zeta = \sum_{\alpha_1=1}^{n_1} \ldots \sum_{\alpha_d=1}^{n_d} L_{\zeta,\alpha_1,\ldots,\alpha_d} X_{\alpha_1,\ldots,\alpha_d}\).

This tensor \(L\) must itself somehow allow for an efficient handling. Similar to Theorem 5.6, each \(L \in \mathbb{R}^{\ell \times n_1 \times \ldots \times n_d}\) can for some \(r_L \in \mathbb{N}^K\) (assumed to be low) be decomposed into lower dimensional components. We therefore define

\[
\mathcal{D}_{r_L}^L := \bigotimes_{v \in V} \mathcal{S}_m^L, \quad m_c^L := \{e^v\}_{e \in E_c} \cup \{\zeta, \ z \} \quad \text{if} \ c \in [d],
\]

otherwise.
The symbols $e^\alpha$, $e \in E$, are additional labels with range $e^\alpha \in [r_L^{(L)}]$, whereas $\zeta \in [\ell]$. The assigned multilinear representation map is
$$
\rho_{r_L}(L)_{\zeta,\alpha_1,\ldots,\alpha_d} := \sum_{e^\alpha, e \in E} \prod_{\mu \in [d]} (L_\mu)_{\zeta,\alpha_\mu,(e^\alpha)_\mu \in E_\mu} \prod_{v \in V \setminus [d]} (L_v)_{(e^\alpha)_v \in E_v}.
$$

For simplicity of notation, as with the representation $N$, we will also denote indices $\zeta$ and $\alpha_\mu$ in nodes $L_v$ with $v > d$. While this is formally compatible as long as $(L_v)_{\zeta,\alpha_\mu,(e^\alpha)_\mu \in E_\mu}, v > d$, is constant in $\zeta$, these indices can likewise be omitted.

Sampling operators for instance can be decomposed for $r_L^{(L)} \equiv 1, J \in K$.

**Example 6.1.** For $d \in \mathbb{N}$, the operator decomposition corresponding to the Tucker graph $K_{\text{Tucker}} (5.3)$ consists of the components $\{L_v\}_{v \in V}$ of sizes $L_\mu \in \mathbb{R}^{\ell \times n_\mu \times r^{(\mu)}}$ and $L_{d+1} \in \mathbb{R}^{r^{(1,d+1)} \times \cdots \times r^{(d,d+1)}}$. The corresponding contraction map $\rho_{r_L}$ is
$$
L_{\zeta,\alpha_1,\ldots,\alpha_d} = \rho_{r_L}(L_1,\ldots,L_d,L_{d+1})_{\zeta,\alpha_1,\ldots,\alpha_d}
$$

for $\alpha_\mu = 1,\ldots,n_\mu$, $\mu = 1,\ldots,d$ and $\zeta = 1,\ldots,\ell$ as visualized in Fig. 2.

In general, when all summations over $\alpha_v, v \in [d]$, are proceeded first, then $\mathcal{L}(X)$ can be efficiently evaluated by means of the tree structure of (cf. Proposition 5.5)

$$
\mathcal{L}(X)_\zeta = \sum_{e^\alpha, e \in E} \prod_{v \in V} (\sum_{\alpha_v} (L_v)_{\zeta,\alpha_\mu,(e^\alpha)_\mu \in E_\mu}(N_v)_{\alpha_\mu,(e^\alpha)_\mu \in E_\mu}), \quad \zeta \in [\ell].
$$

Note that we have here again made use of the redundant additional indices $\zeta$ and $\alpha_\mu$ for $v > d$. Likewise, the composition of $\mathcal{L}$ and $\mathcal{N}_{\neq c}$ can be proceeded efficiently.

### 6.4. Equivalent low rank weights

The switching between each complementary weights introduced in Section 3.2 has the following motivation.

**Lemma 6.2.** Let $c \in V$ and $S_c$ be as in Lemma 5.10, and let $N$ be a representation for which $Y^{(J)}$, $J \in K^{S_c}$, are orthonormal (cf. Theorem 5.11). Then the update $X_{\gamma,c,W}$ as defined in (6.2) for the rank $n_J$ matrices $W^{(J)} = W^{(J)}_{\gamma,N,c} = (X^{(J)}(X^{(J)})^T + \gamma I)^{-1}, J \in K^{S_c}$, $X = \tau_r(N)$, is the same as for the rank $r^{(J)}$ matrices $W^{(J)} = W^{(J)}_{\gamma,N,c} := Y^{(J)}(H^{(J)} + \gamma I)^{-1}(Y^{(J)})^T, \quad H^{(J)} := Z^{(J)}(Z^{(J)})^T, \quad J \in K^{S_c}$.

**Proof.** It suffices to show that for every $\tilde{N}_c \in S_m$, we have
$$
(W_{\gamma,N,c}^{(J)})^{1/2}N_{\neq c}(\tilde{N}_c)^{[J]} = (W_{\gamma,N,c}^{(J)})^{1/2}N_{\neq c}(\tilde{N}_c)^{[J]}, \quad J \in K^{S_c}.
$$
Let the orthonormal matrix \( U^{J,\perp} \in \mathbb{R}^{n_J \times n_{J^\perp}} \) span the orthogonal complement of the \( r(J) \) dimensional space range(\( Y^{(J)} \)). Then
\[
(W_{s,X}^{(J)})^{1/2} = (W_{s,X}^{(J),\perp} + \gamma^{-1} U^{J,\perp} (U^{J,\perp})^T)^{1/2} = (W_{s,X}^{(J),\perp})^{1/2} + \gamma^{-1/2} U^{J,\perp} (U^{J,\perp})^T.
\]

It thus remains to show that range(\( N_{p}\mathcal{C}(\tilde{N}_c)^{(J)} \)) \( \perp \) range(\( U^{J,\perp} \)) for all \( \tilde{N}_c \). As by construction of \( S_c \), the matrix \( Y^{(J)} \) does not depend on the vertex \( c \in V \), we have
\[
\text{range}(N_{p}\mathcal{C}(\tilde{N}_c)^{(J)}) \subseteq \text{range}(Y^{(J)}).
\]

\( \square \)

### 6.5. Path evaluations

Lemma 6.2 allows for a further, significant simplification in the evaluation of \((N_c)^{(J)}_{\mathcal{C}:W}\) as defined in (6.3). Let in the following \( c \in V \) and \( J \in \mathcal{K}^S \) be fixed. Further, let \( v_{c,j} \in V \setminus \{c\} \) be the uniquely determined vertex with \( J(c,v_{c,j}) \cap [d] = J \) (cf. Section 5.2), as in Section 5.7. Without explicit indication of the dependence on the above, we denote (cf. (5.2) and (5.4))
\[
p := c \mapsto v_{c,j} \subseteq V \setminus \{c, v_{c,j}\}, \quad E_Y := \partial E_{\{c\}} \cup p = E_c \setminus \{c_1\} \cup \partial E_p.
\]

For empty \( p \), we set \( p_1 = v_{c,j} \) and \( p_{-1} = c \) for convenience. We further define the edges \( e_1 = \{c, p_1\} \notin E_Y \) and \( e = \{p_{-1}, v_{c,j}\} \in E_Y \), such that \( \hat{J} = J_e \).

**Theorem 6.3** (cf. Section SM5.2). Let \( c \in V \) and \( S_c \) be as in Lemma 5.10, and let \( N \) be a representation for which \( Y^{(J)} \), \( J \in \mathcal{K}^S_c \), are orthonormal (as in Theorem 5.11). Further, let the operator \( N_{p}\mathcal{C} \circ W^{(J)} \circ N_{p} : \mathcal{H}_m \to \mathcal{H}_m \) be described by the matrix \( A^{(J)} \in \mathcal{H}_m \otimes \mathcal{H}_m \). Then (cf. Fig. 3)
\[
A^{(J)}_{n \times n_{(\mathcal{C},\mathcal{J})}} = \delta_{n,n_{(\mathcal{C},\mathcal{J})}}(\prod_{\alpha \in E_c \setminus \{\alpha_1\}} \delta_{(\beta',\gamma)_{\alpha}}) \cdot M^{(J)}_{(\alpha',\beta')_{\gamma}}(\prod_{\beta \in E_p \setminus \{\beta_1\}} \delta_{(\beta',\gamma)_{\alpha}}),
\]
where each \( \delta_{\alpha',\gamma} \in \{0,1\} \) is a Kronecker delta, as well as
\[
M^{(J)}_{(\alpha',\beta')_{\gamma}} = \sum_{\alpha \in \text{v.e.p.}} (\prod_{\beta \in \text{v.e.p.}} \delta_{(\beta',\gamma)_{\alpha}}) \quad \left( \prod_{\alpha \in \text{E深层次}} (N_{e})_{\alpha \leftarrow (\beta',\gamma)_{\alpha}} (N_{e})_{\alpha \leftarrow (\beta',\gamma)_{\alpha}} \right) (H^{(J)} + \gamma I)^{-1}_{(\beta',\gamma)}
\]
and further (cf. Fig. 4)
\[
H^{(J)}_{(\beta',\beta)} = \sum_{\alpha \in \text{E深层次}, \beta \in \text{v.e.p.}} (\prod_{\alpha \in \text{E深层次}} \delta_{(\beta',\gamma)_{\alpha}}) \quad \left( \prod_{\alpha \in \text{E深层次}} (N_{e})_{\alpha \leftarrow (\beta',\gamma)_{\alpha}} (N_{e})_{\alpha \leftarrow (\beta',\gamma)_{\alpha}} \right),
\]
for each \( \alpha \in [n_u], v \in V, \) and \( \beta', \beta \in \alpha_{(J_e)}, e \in E \).

**Proof.** See Figs. 3 and 4. For the rigorous, though exceedingly technical proof, see Section SM4. A more elegant version can be found in Section SM5.2.

\( \square \)

The formula for \( A^{(J)} \) simplifies whenever \( e \in E_c \) as follows.
Corollary 6.4. In Theorem 6.3, if \( \hat{e} = e_1 = \{c,v\} \) for \( v \in \text{neigh}(c) \), then \( p = \emptyset \). Thus, we have \( M^{(J)} = (H^{(J)} + \gamma I)^{-1} \) and
\[
H^{(j)}_{\beta e', \beta e} = \sum_{\alpha_e \in [n_e], \beta^e : e \in E_c \setminus (\hat{e})} (N_c)_{\alpha_e, (\beta^e)_{e \in E_c \setminus (\hat{e})} \cdot \beta e'} \cdot (N_c)_{\alpha_e, (\beta^e)_{e \in E_c}}
\]
for \( \alpha_e \in [n_e] \), and \( \beta e', \beta^e \in \{ \beta^{(J)} \}, e \in E_c \).

6.6. Branch evaluations. As described in the following, each expression in the update formula of \( (N_e)_{\gamma^e, W} \) (cf. (6.3)) can be rewritten, such that reevaluations of identical terms are avoided. This is particularly useful (Proposition 6.5) for the first measurement related summand \( (L \circ N_{\rho e})^* \circ L \circ N_{\rho e} : \mathcal{H}_m \to \mathcal{H}_m \) and the righthand side \( (L \circ N_{\rho e})^* : \mathbb{R}^r \to \mathcal{H}_m \), since only few branch-wise evaluations change after each micro-step during a sweep. While this is not true for the weight related terms or the sum of such, the computational complexity may (depending on \( c \in V \), \( K \) and \( d \)) still be reduced through the recursive, branch-wise evaluation of the entire term \( \sum_{J \in K \cap c} N_{\rho e} \circ \mathcal{W}^{(j)} \circ N_{\rho e} \) (Propositions 6.6 and 6.7).
**Proposition 6.5.** Let \( c \in V \) and each \( J_e \in K^{S_c} \), \( e \in E \). Further, let \( \mathcal{L} \circ \mathcal{N}_{\neq c} : \mathcal{H}_{m_c} \to \mathbb{R}^t \) be represented by the tensor \( F_c \in \mathbb{R}^{[t]} \otimes \mathcal{H}_{m_c} \). Then

\[
(F_c)_{\xi, \alpha_e, \{e^*\}_{e \in E_c}} = \sum_{e^* : e \in E_c} \left( L_e \right)_{\xi, \alpha_e, \{e^*\}_{e \in E_c}} \prod_{v \in \text{neigh}(c)} S_{\xi, \beta_e, \{e^*\}_{e \in E_v}}^{(J_v, \{e\})},
\]

for \( \xi \in [t] \) (not being contracted), as well as

\[
S_{\xi, \beta_e, \{e^*\}_{e \in E_v}}^{(J_v, \{e\})} = \sum_{e^*, \beta^* : e \in E_v \setminus \{\ell\}} \left( L_v \right)_{\xi, \alpha_e, \{e^*\}_{e \in E_v}} \left( N_v \right)_{\alpha_e, \{\beta^*\}_{e \in E_v}} \prod_{b \in \text{desc}_v(v), \{\beta^*\}_{e \in E_v}} S_{\xi, \beta_e, \{e^*\}_{e \in E_v}}^{(J_v, \{e\})},
\]

for \( e = \{\text{pred}_c(v), v\}, v \in V \setminus \{c\} \) and \( \xi \in [t] \) (not being contracted).

**Proof.** See Section SM5.3.

The paths appearing in the evaluation of \( A^{(J)} \in \mathcal{H}_{m_c} \otimes \mathcal{H}_{m_e} \), representing \( \mathcal{N}_{\neq c} \circ \mathcal{W}^{(J)} \circ \mathcal{N}_{\neq c} : \mathcal{H}_{m_c} \to \mathcal{H}_{m_e}, \ J \in K^{S_c} \) (cf. Theorem 6.3) naturally overlap. In the evaluation of \( A := \sum_{J \in K} A^{(J)} \) (cf. (6.3)), this can be utilized.

**Proposition 6.6.** Let \( c \in V \) and each \( J_e \in K^{S_c} \), \( e \in E \). It is

\[
\sum_{J \in K} A^{(J)}_{\alpha_e', \{\beta^*\}_{e \in E_v}; \alpha_e, \{\beta^*\}_{e \in E_v}, \alpha_e} = \delta_{\alpha_e', \alpha_e} \sum_{e \in \text{neigh}(c)} \left( \prod_{e \in \{\ell, c\}} \delta_{\beta_e', \beta_e} \right) B^{(J_{\ell, c})}_{\beta_e, \{\beta^*\}_{e \in E_v}, \alpha_e} \cdot B^{(J_{\ell, c})}_{\beta_e, \{\beta^*\}_{e \in E_v}, \alpha_e},
\]

with \( B^{(J_{\ell, c})} = (H^{(J_e)} + \gamma I)^{-1} + \sum_{b \in \text{desc}_v(v)} B^{(J_{b, \ell})}_{b, \{\beta^*\}_{e \in E_v}, \alpha_e} \), \( e = \{\text{pred}_c(v), v\}, v \in V \setminus \{c\} \), as well as, for \( b \in \text{desc}_v(v) \),

\[
B^{(J_{b, \ell})}_{b, \{\beta^*\}_{e \in E_v}, \alpha_e} = \sum_{\beta^*: e \in E_v \setminus \{\ell\}, \{\beta^*\}_{e \in E_v} \neq \{\beta_{\ell}^*\}_{e \in E_v}} \left( N_v \right)_{\alpha_e, \beta_{\ell}^*, \{\beta^*\}_{e \in E_v \setminus \{\ell\}}} \cdot B^{(J_{b, \ell})}_{b, \{\beta_{\ell}^*\}_{e \in E_v}, \alpha_e} \cdot \left( N_v \right)_{\alpha_e, \{\beta^*\}_{e \in E_v}.}
\]

**Proof.** See Section SM5.3.

Due to the recursive structure in Proposition 6.6, the evaluation is to be proceeded in order leaves to root. In turn, also the matrices \( H^{(J_e)} \), \( \ell, \epsilon \in E \), (cf. (6.8)) can be simplified, but in the opposing root to leaves order. The starting points for this recursion are given by Corollary 6.4.

**Proposition 6.7.** Let \( c \in V \) and each \( J_e \in K^{S_c} \), \( e \in E \). For \( e = \{\text{pred}_c(v), v\}, v \in V \setminus \{c\} \), and \( b \in \text{desc}_v(v) \), it is

\[
H^{(J_{b, \ell})}_{b, \{\beta_{\ell}^*\}_{e \in E_v}, \alpha_e} = \sum_{\beta^*: e \in E_v \setminus \{\ell\}} \left( N_v \right)_{\alpha_e, \beta_{\ell}^*, \{\beta^*\}_{e \in E_v \setminus \{\ell\}}} \cdot \left( N_v \right)_{\alpha_e, \{\beta^*\}_{e \in E_v}.}
\]

**Proof.** See Section SM5.3.

### 7. Numerical Experiments

The following Sections 7.1 to 7.4 specify terminology and configurations referred to in the subsequent experiments SM1 and 7.1 to 7.4 in Sections SM1, 7.7 and 7.9. The presentation of results is further laid out in Section 7.5. For simplicity, the mode sizes \( \{n^\mu\}_{\mu \in [\ell]} \) are chosen uniformly as \( \pi \in \mathbb{N} \) in all experiments. For the corresponding MATLAB code, please contact the author.
7.1. Reference solutions, measurements vectors and family $\mathcal{K}$. Each measurement vector is constructed via a (not necessarily sought for) reference solution with ranks $r_{(rs)} \in \mathbb{N}^\mathcal{K}$, which in turn relies on a randomly generated representation, 
\[
y = \mathcal{L}(X_{(rs)}) \in \mathbb{R}^\ell, \quad X_{(rs)} = \tau_{(rs)}(N_{(rs)}) \in \mathcal{L}^{-1}(y) \cap V^\mathcal{K}_{\leq r_{(rs)}}.
\]
All entries of the components $\{N_v^{(rs)}\}_{v \in \mathcal{V}^{(rs)}}$ are assigned independently, normally distributed entries. For simplicity $^7$ and to limit the amount of randomness, we also choose the components $\{\rho_{(rs)}^{(j)}\}_{J \in \mathcal{K}}$ uniformly, as $\tau_{(rs)} \in \mathbb{N}$. We distinguish between four different types.

**Tucker format.** With $\mathcal{K} = \mathcal{K}_{\text{Tucker}} = \{\{1\}, \ldots, \{d\}\}$, the components of the representation $\{N_v^{(rs)}\}_{v \in \mathcal{V}^{(rs)}}$ follow the scheme in Example 5.5.

**Balanced, binary hierarchical Tucker format (bbHT).** A balanced, binary hierarchical Tucker format can be defined by the property of (cf. Section 5.6) and to minimize the maximal distance of any two vertices $v, w \in [d]$ within $G_K$ (that is, the depth of the rooted tree, cf. [15]).

**Exponentially declining singular values.** Firstly, a bbHT representation as defined above is generated. As second step, all singular values $\sigma^{(J)}$, $J \in \mathcal{K}$, are manipulated such that they decline exponentially. In explicit, for a constant $\gamma_{(expfac)} \in (0, 1)$, it is $\sigma^{(J)} \approx \max(\sigma_{\min}, \gamma_{(expfac)})$, $i = 1, \ldots, r_{(rs)}$, $J \in \mathcal{K}$, where each $x$ is an independent random, normally distributed value and $\sigma_{\min} > \varepsilon > 0$ (cf. Section 7.4) is a lower bound. We denote such reference solutions by the abbreviated $\text{exp.dec.bbHT}$.

**Canonical polyadic (CP) decomposition.** For $\tau_{(rs)} \in \mathbb{N}$, the reference solution does here not rely on $\mathcal{K}$, but is generated as sum of $\tau_{(rs)}$ elementary tensors, $X_{(rs)}^{\alpha_1, \ldots, \alpha_d} := \tau_{(rs)}(\phi^{(rs)}) = \sum_{\alpha_1, \ldots, \alpha_d}^{\tau_{(rs)}}(\phi^{(rs)}_{\alpha_1, \ldots, \alpha_d}),$ where $\phi^{(rs)}_{\mu} \in \mathbb{R}^{[n_{\mu}] \times [\tau_{(rs)}]}$, for $\mu = 1, \ldots, d$. The corresponding graph is a hypertree, and the set $\{\tau_{(rs)}\}$ of at most rank $\tau_{(rs)}$ tensors is a semi-algebraic subset of $V^\mathcal{K}_{\leq r_{(rs)}}$ (cf. (1.7)) for the in that case defined, non-hierarchical family $\mathcal{K}_{\max} := \{J \subseteq [d] \mid J \neq \emptyset\}$, given $r^{(J)}_{(rs)} \equiv \tau_{(rs)}$, $J \in \mathcal{K}_{\max}$.

7.2. Operators. We consider three types of operators $\mathcal{L}$, where in each case $\mathcal{L}(X) := L \text{vec}(X)$ is the tensor on the tensor $L \in \mathbb{R}^{t \times n_1 \ldots n_d}$.

**(Full) Gaussian operator.** With a Gaussian operator, we refer to a randomly generated tensor $L \in \mathbb{R}^{t \times n_1 \ldots n_d}$ with independent, normally distributed entries.

**Gaussian low rank operator.** For (low) uniform ranks $r^{(j)}_L \equiv r_L \in \mathbb{N}$, $J \in \mathcal{K}$, the operator is defined through the representation of $L := \rho_{(rs)}(\{L_v\}_{v \in \mathcal{V}^{(rs)}})$ (cf. Section 6.3). Each component therein are assigned independent, normally distributed entries.

**Random sampling operator.** As sampling operator, we denote $\mathcal{L}(X) := \{X_{p_i}\}_{i=1}^\ell$, for uniformly randomly drawn indices $\{p_1, \ldots, p_\ell\} \subset \times_{\mu=1}^d [n_{\mu}]$. Note that sampling operators can trivially be decomposed, for $r^{(J)}_L = 1$, $J \in \mathcal{K}$.

7.3. Solution methods. Based on a sufficiently large starting value $\gamma^{(0)} > 0$, we choose $\gamma^{(i)} = \nu \gamma^{(i-1)}$, where $\nu < 1$ remains constant throughout each single run of an algorithm. We consider the following types of optimization.

**Full, image based (IRLS-0K).** As in (1.4), the full tensor is optimized based on the (literally interpreted) image update formula (3.3) without further modification (Algorithm 1 for $\mathcal{S}_i \equiv \emptyset$, $i \in \mathbb{N}_0$). When instability threatens to occur, the equivalent kernel based update (3.6) for $X_0 = X^{(0)}$ is applied, with $X^{(0)}$ as in Corollary 3.2.
Full, relaxed. The relaxed constraints described in Section 4.1 are utilized, but without subspace restrictions or weight switching (Algorithm 2 with $T_i \equiv \mathcal{L}^{-1}(y)$, $S_i \equiv \emptyset$, $i \in \mathbb{N}_0$). In this case, the residual $\|\mathcal{L}(X) - y\|$ is expected to converge to 0 parallel to the decline of $\gamma$, but this is not guaranteed.

Alternating (AIRLS-0K). We apply the representation based, necessarily relaxed, alternating optimization (Algorithm 2 for $T_i = T_i(\tau_i(N^{(i)}))$ and $S_i = S_c$, $i \in \mathbb{N}_0$) further discussed in Section 6. The update formulas for the single components make use of the branch-wise evaluations as derived in Section 6.6. Whether the (maximal) ranks $r \in \mathbb{N}^K$ of the iterate, that is, the sizes of $\{N_v\}_{v \in V}$, are fixed or adapted, as well as the potential use of other heuristics laid out in Section SM6, is specified in the respective experiments.

Neigh. The same algorithm as aboves AIRLS-0K is applied, but in each update of the node $N_c$, $c \in V$, only weights corresponding to $J_c \in K^{c \epsilon}$, $c \in V$, are included (cf. Corollary 6.4) in order to reduce the computational complexity. This reduction of paths yields the variant closest to our priorly introduced algorithm SALSA, and in particular the minimal number of weights in each the update of $N_c$, $c \in V$, for which the rank adaption stability property, as further introduced in [16], still holds true.

Plain ALS without reweighting. In one instance in Experiment 7.3, we also compare to the plain alternating least squares (ALS) residual minimization ((6.2) for $\omega = \gamma = 0$) for fixed ranks $r = r_{\text{rs}} \in \mathbb{N}^K$. This algorithm is thus granted additional, in practice generally unavailable information and does not adapt ranks.

7.4. Experimental setup and evaluation. In order to evaluate each output $X^{(\text{alg})}$, we compare its non-neglectable singular values to those of $X^{(\text{rs})}$. We define $\det_{n,x,\gamma,\epsilon}(X) := \prod_{J \in K} \det_{n,x,\gamma,\epsilon}(X^{(J)})$, $X \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, where the matrix version is as in [26] given by

$$
\det^2_{m,\gamma,\epsilon}(A) := \gamma^{m - \text{rank}_x(A)} \prod_{i=1}^{\text{rank}_x(A)} (\sigma_i(A)^2 + \gamma),
$$

for $\text{rank}_x(A) := \max\{i \in [m] \mid \sigma_i(A) > \epsilon \cdot \|A\|_F\}$. Therein, we choose $\epsilon := 10^{-6}$. We firstly examine the residual norm $\|\mathcal{L}(X^{(\text{rs})}) - y\|_F$, secondly compare the approximate ranks, and lastly compare the products of singular values. The latter two aspects are reflected by the limit of the quotient $Q_{\epsilon}(X^{(\text{alg})}, X^{(\text{rs})}) := \lim_{\gamma \to 0} \frac{\det^2_{n,\gamma,\epsilon}(X^{(\text{alg})})}{\det^2_{n,\gamma,\epsilon}(X^{(\text{rs})})} \in [0, 0.98] \cup (0.98, 1.005) \cup [1.005, \infty]$.

The three intervals are related to the categorization into improvements, successes or the two types of failures as outlined below, where the limits 0 or $\infty$ are reached if and only if $\sum_{J \in K} \text{rank}_x((X^{(\text{alg})})^{[J]})$ and $\sum_{J \in K} \text{rank}_x((X^{(\text{rs})})^{[J]})$ differ.

Post iteration. In order to avoid misjudgment, in cases where the tensor $X^{(\text{alg})}$ may be an improving solution (though that seldomly happens here), we apply a post iteration analogous to the one discussed in the matrix case [26] in order to allow the parameter $\epsilon$ to be reduced to machine precision.

Details of comparison. As in [26], if $\|\mathcal{L}(X^{(\text{alg})}) - y\| > 10^{-6}\|y\|$ or if for the quotient, it holds $Q_{\epsilon}(X^{(\text{alg})}, X^{(\text{rs})}) = \infty$, then the result is considered a strong failure. If $\|\mathcal{L}(X^{(\text{alg})}) - y\| \leq 10^{-6}\|y\|$, then on the one hand we refer to $1.005 \leq Q_{\epsilon}(X^{(\text{alg})}, X^{(\text{rs})}) < \infty$ as weak failure. On the other, for $0.98 < Q_{\epsilon}(X^{(\text{alg})}, X^{(\text{rs})}) < 1.005$, we consider the result successful, while for $Q_{\epsilon}(X^{(\text{alg})}, X^{(\text{rs})}) \leq 0.98$, we say the result is an improvement, subject to the consideration above.
7.5. Presentation of results. Each but experiments 7.1 and 7.3 is reflected upon in three different ways as summarized in Table SM2.

ASRM/recovery tables. For each instance, we list the percentual numbers of ASRM improvements, successes or fails as defined in Section 7.4. Successes are further distinguished regarding recoveries, whether \( \|X'(\nu) - X(\nu')\|_F \leq 10^{-4}\|X(\nu')\|_F \). In near all cases where this is fulfilled, the relative residual even falls below \( 10^{-6} \) (see Section SM2), in which case the algorithm stops automatically. Note that both improvements as well as fails with respect to ASRM naturally nearly exclude recoveries with accuracy \( 10^{-4} \), and always so for \( 10^{-6} \).

ASRM/recovery figures. More distinguished visualizations of the results underlying the above mentioned tables can be found in Section SM2 as described therein.

\( \gamma \)-decline sensitivity. A depiction of results regarding the sensitivity analysis outlined in Section 7.4 is covered in Section SM2 as well.

7.6. Observing the theoretical phase transition for generic recoveries.

Experiment 7.1. For \( d = 4, \pi = 5 \) and \( \tau_{(rs)} = 3 \), we consider the ASRM-\( K_{bbHT} \) problem based on Gaussian measurements for reference solutions given via \( bbHT \) representations for \( \ell \in \{68, 69, 70\} \). The solution method in both cases utilizes full, image based updates (cf. Section 7.1). Each constellation is repeated 100 times, for a comparatively large value of \( k_{\text{max}} = 10 \). The results are covered in Table 1.

The dimension of the given \( bbHT \) variety is \( \text{dim}(V^r_{\leq \ell (rs)}) = 4m\tau_{(rs)} + 2\tau_{(rs)}^3 - 5\tau_{(rs)}^2 = 69 \) (cf. Lemma 5.7). The value \( \ell \equiv \text{dim}(V^r_{\leq \ell (rs)}) + 1 \) (which here is \( \ell = 70 \)) in turn provides the minimal sufficient number of generic\(^9\) measurements (thus not including sampling) to provide \( \mathcal{L}^{-1}(\mathcal{L}(X(\nu))) \cap V^r_{\leq \ell (rs)} = \{X(\nu)\} \) for generic \( X(\nu) \in V^r_{\leq \ell (rs)} \), as more generally proven in [4]. We can indeed observe (see Table 1) that for \( \ell = 69 \), multiple solutions are found as verified through the post iteration process up to machine accuracy. For the value \( \ell = 70 \) in turn, no duplicate solutions seem to exist. The one improving solution as well as the two weak failures are not the reference solution, though in fact neither within \( V^r_{\leq \ell} \) for \( r = \tau_{(rs)} \) but \( r = \tilde{r} \), \( \tilde{r}(1.2) = 3 \), \( \tilde{r}(1) \), \( \tilde{r}(4) \) = (2, 2, 4, 4). From the perspective of a dimension minimization (cf. (2.1)) in turn, not even the improving result would be preferable as \( \text{dim}(V^r_{\leq \ell}) = 71 \) (\( V^r_{\leq \ell} \nsubseteq V^r_{\leq \ell (rs)} \)).

7.7. Affine sum-of-ranks minimization.

Experiment 7.2. For \( d = 4, \pi = 5 \) and \( \tau_{(rs)} = 3 \), we consider the ASRM-\( K \) problem based on samples or Gaussian measurements and reference solutions given via \( bbHT \) representations for \( \ell \in \{83, 111, 138\} \) and \( K = K_{bbHT} \) or by \( CP \) decompositions for \( \ell \in \{62, 82, 102\} \) and \( K = K_{\text{max}} \). The solution method in both cases utilizes full, image based updates based on the respective families \( K \) (cf.

---

\(^8\)Needless to say, this is the only point at which the reference solution itself is used within the algorithm, and only done in order to save a considerable amount of unnecessary computation time.

\(^9\)To be more precise, generic in that context is an algebraic property that is stronger than the ones that stem from analysis or probability theory, but roughly similar.
Section 7.1). Each constellation is repeated 100 times, for \( k_{\text{max}} = 8 \). The results are covered in Table 2 and Figs. SM3 and SM4.

| instance | ASRM-\( K_{\text{K-HIT}} \) / max. recovery | \( Q_x \in [0, 0.98] \) | \( Q_x \in (0.98, 1.005) \) | \( Q_x \in [1.005, \infty) \) | \( Q_x = \infty \) |
|----------|---------------------------------|----------------|----------------|----------------|----------------|
| \( c_{\text{mf}} = 1.2 \) | \begin{itemize}  
  \item gaussian, HT, \( t = 83 \)  
  \item gaussian, CP, \( t = 62 \)  
  \item sampling, HT, \( t = 83 \)  
  \item sampling, CP, \( t = 62 \)  
\end{itemize} | 0 | 0 | + | 96 | 0 | 4 |

| \( c_{\text{mf}} = 1.6 \) | \begin{itemize}  
  \item gaussian, HT, \( t = 111 \)  
  \item gaussian, CP, \( t = 82 \)  
  \item sampling, HT, \( t = 111 \)  
  \item sampling, CP, \( t = 82 \)  
\end{itemize} | 0 | 0 | + | 100 | 0 | 0 |

| \( c_{\text{mf}} = 2.0 \) | \begin{itemize}  
  \item gaussian, HT, \( t = 138 \)  
  \item gaussian, CP, \( t = 102 \)  
  \item sampling, HT, \( t = 138 \)  
  \item sampling, CP, \( t = 102 \)  
\end{itemize} | 0 | 0 | + | 100 | 0 | 0 |

Table 2. IRLS-0K (full, image method, \( d = 4 \), \( \pi = 5 \), \( \tau_{\text{rs}} = 3 \)) – table as specified in Section 7.5 for Experiment 7.2 (see Fig. SM4 for more details)

The dimension or even the more particular structure of the variety \( V_{\leq \tau_{\text{rs}}}^{K_{\text{max}}} \), for \( K_{\text{max}} = \{ J \subseteq [d] \mid J \neq \emptyset \} \), \( d \geq 4 \), as applied in the CP case (cf. Section 7.1), is unknown to the best of our knowledge. While real tensors of at most rank \( \tau_{\text{rs}} \) do not form varieties, complex ones with at most this border rank do, here with a dimension of \( \dim(V_{\leq \tau_{\text{rs}}}^{K_{\text{max}}} \subset \subset) = \tau_{\text{rs}}(d(\pi - 1) + 1) = 51 \) (cf. [3, 6, 33]). Though we assume this dimension to be lower or equal to the one for \( K_{\text{max}} \), we take this smaller value as reference. In that sense, the considered values \( \ell \) are each (rounded) multiples \( c_{\text{mf}} \in \{1.2, 1.6, 2\} \). To our surprise, if successful, the CP reference solution is (near perfectly) recovered even for \( \ell = 62 \), considering that this value is smaller than 69 \( = \dim(V_{\leq \tau_{\text{rs}}}^{K_{\text{max}}} \subset \subset) \) for every exhaustive hierarchical family \( K_{\text{K-HIT}} \). One possible explanation would be that \( \dim(V_{\leq \tau_{\text{rs}}}^{K_{\text{max}}} \subset \subset) \) is lower or equal to 61, but further investigation remains subject to future work. While we can not, as theory provides, expect generic completions in case of sampling problems, the failures with respect to ASRM are subject of IRLS-0K itself. In particular, slower rates of decline \( \nu \) (cf. Fig. SM3) may be required, and allow for better results for both sampling and Gaussian measurements as suggested by Fig. SM1. Though already for \( c_{\text{mf}} = 1.2 \), the rate of decline seems to suffice.

7.8. Alternating, affine sum-of-ranks minimization.

Experiment 7.3. For \( d = 4 \), \( \pi = 5 \), \( \tau_{\text{rs}} = 3 \) and \( \ell \in \{126, 168, 210\} \) we consider the ASRM-\( K_{\text{Tucker}} \) problem based on samples or rank \( \tau_L = 1 \) Gaussian measurements for reference solutions given through Tucker representations. We compare the following four solution methods:

(a) full, image based

(c) alternating, based on fixed, maximally feasible ranks \( r(J) = 5 \), \( J \in K_{\text{Tucker}} \)

(d) alternating, with adaptive ranks \( r(J) \in [5] \), \( J \in K_{\text{Tucker}} \) (cf. Section SM6.3)

(e) plain, ALS without reweighting, based on the, a-priori provided, fixed ranks \( r(J) = \tau_{\text{rs}} = 3 \), \( J \in K_{\text{Tucker}} \).
A fixed rate $\nu = 1.002^{-1}$ of decline is used (applicable to the first three methods). Each constellation is repeated 100 times, for which the results are covered in Table 3 and Fig. SM7.

| instance | $\nu \in [0.98, 1.005]$ | $\nu \in (0.98, 1.005)$ | $\nu \in [1.005, \infty)$ | $\nu = \infty$ |
|-----------|-----------------|-----------------|-----------------|-------------|
| *gaussian ($\nu_L = 1$): (a) full* | no | no | yes | 0 |
| * (c) alt | 0 | 0 + 79 | 0 | 21 |
| * (d) alt adapt | 0 | 0 + 78 | 0 | 22 |
| * (e) plain ALS | 0 | 0 + 100 | 0 | 23 |

Table 3. (A)IRLS-0K$_{bbHT}$ ($T = 4, \nu = 5, \nu_{rs} = 3$) – table as specified in Section 7.5 for Experiment 7.3 (see Fig. SM7 for more details)

The degrees of freedom within a Tucker decompositions for $d = 4$ in this setting is $\dim(V^{\nu = 4}_{\leq \nu_L}) = \nu^4_{rs} + 4 \nu^2_{rs} - 4 \nu^2_{rs} = 105$ (cf. Lemma 5.7). The number of measurements $\ell$ are (rounded) multiples $c_{nr} \in \{1, 2, 1.6, 2\}$ of such. As in Experiment SM1, there is nearly no difference between the version using fixed ranks or adaptive ranks, but both instances are slightly worse than the full version using unrelaxed constraints (note that here, these methods use the same, fixed rate of decay $\nu = 1.002^{-1}$). Plain alternating least squares on the other hand (even though only in that case, the ranks of the reference solution are provided) is significantly worse than the other methods, also for larger numbers of measurements.

7.9. Large scale, alternating ASRM.

Experiment 7.4. For $d = 8, \nu = 20, \nu_{rs} = 5$ and $\ell \in \{6500, 13000, 19500, 26000\}$, we consider the ASRM-$K_{bbHT}$ problem based on samples, rank $\nu_L = 1$ or rank $\nu_L = 2$ Gaussian measurements for reference solutions given via $bbHT$ representations with exponentially declining singular values, $s(\text{expfac}) = \frac{1}{3}$. For Gaussian measurements, we also consider unmodified singular values. As solution method, we apply alternating optimization with explicit rank adaption (limited only by $r(J) \leq 8, J \in K_{bbHT}$) as well as the applicable heuristics laid out in Section SM6. The maximal length of paths is either unrestricted, or limited to neighbors. Each constellation is repeated 100 times, for $k_{\text{max}} = 5$. The results are covered in Tables 4 and 5 and Figs. SM8 to SM11.

The degrees of freedom within 8-dimensional $bbHT$ decompositions in this setting is $\dim(V^{\nu = 4}_{\leq \nu_L}) = 8 \nu^4_{rs} + 6 \nu^2_{rs} - 3 \nu^2_{rs} = 1225$ (cf. Lemma 5.7), while $\ell = 6500$ constitutes a fraction of about $2.5 \cdot 10^{-7}$ of the total size $\nu^4 = 2.56 \cdot 10^{10}$ of the tensor. Due to the long runtime for values $k > 5$, it yet remains speculation whether the restriction of paths to neighboring nodes does result in a loss of approximation quality or, as in other cases, rather a need for a lower parameter $\nu$ (cf. Figs. SM8 and SM10). The same might hold true for the completion problem considered here.
Conclusions and outlook

We have shown that despite subtle differences, the overall structure of the log-det approach towards ARM can be generalized to the ASRM tensor setting. The global convergence of minimizers of the log-det sum-of-ranks function can likewise be concluded via the priorly applied nested minimization scheme. Even subject to the additionally considered switching between complementary subsets in $K$, the IRLS-$0K$ algorithm inherits analogous local convergence properties, in particular with respect to the decline of the regularization parameter $\gamma \downarrow 0$. Thereafter, we have laid out that despite the relaxation of the affine constraint, as well as the iterative restriction to admissible subspaces, IRLS-$0K$ remains faithful to a monotone minimization of the corresponding objective function. In particular, these modifications allow a tree tensor network based, alternating evaluation AIRLS-$0K$, with a non-exponential, low computational complexity based on branch-wise evaluations.

Table 4. (A)IRLS-$0K_{\text{BHT}}$ (alternating, with heuristics, $d = 8$, $\Pi = 20$, $\tau_{m} = 5$) – table as specified in Section 7.5 for Experiment 7.4 (see Fig. SM9 for more details)

| instance | ASRM-$\text{BHT}^{-}$ recovery: $Q_{e} \in [0,0.98]$ | $Q_{e} \in (0.98,1.005)$ | $Q_{e} \in [1.005,\infty)$ | $Q_{e} = \infty$ |
|----------|-------------------------------------------------|-------------------|-------------------|-------------------|
| $t = 13000$ | | | | |
| | * gaussian ($r_{L} = 1$), neigh | 0 | 0 | 0 | 100 |
| | * gaussian ($r_{L} = 1$) | 0 | 0 | 0 | 100 |
| | * gaussian ($r_{L} = 2$) | 0 | 0 | 0 | 100 |
| | * gaussian ($r_{L} = 2$), neigh | 0 | 0 | 0 | 100 |
| $t = 19500$ | | | | |
| | * gaussian ($r_{L} = 1$), neigh | 0 | 0 | 0 | 84 |
| | * gaussian ($r_{L} = 1$) | 0 | 0 | 0 | 99 |
| | * gaussian ($r_{L} = 2$) | 0 | 0 | 0 | 68 |
| | * gaussian ($r_{L} = 2$), neigh | 0 | 0 | 0 | 94 |
| $t = 26000$ | | | | |
| | * gaussian ($r_{L} = 1$), neigh | 0 | 0 | 0 | 89 |
| | * gaussian ($r_{L} = 1$) | 0 | 0 | 0 | 46 |
| | * gaussian ($r_{L} = 2$) | 0 | 0 | 0 | 12 |
| | * gaussian ($r_{L} = 2$), neigh | 0 | 0 | 0 | 27 |

Table 5. (A)IRLS-$0K_{\text{BHT}}$ (alternating, with heuristics, $s = 1/3$, $d = 8$, $\Pi = 20$, $\tau_{m} = 5$) – table as specified in Section 7.5 for Experiment 7.4 (see Fig. SM11 for more details)

| instance | ASRM-$\text{BHT}^{-}$ recovery: $Q_{e} \in [0,0.98]$ | $Q_{e} \in (0.98,1.005)$ | $Q_{e} \in [1.005,\infty)$ | $Q_{e} = \infty$ |
|----------|-------------------------------------------------|-------------------|-------------------|-------------------|
| $t = 6500$ | | | | |
| | * gaussian ($r_{L} = 1$), neigh | 0 | 0 | 0 | 49 |
| | * gaussian ($r_{L} = 1$) | 0 | 0 | 0 | 91 |
| | * gaussian ($r_{L} = 2$) | 0 | 0 | 0 | 36 |
| | * gaussian ($r_{L} = 2$), neigh | 0 | 0 | 0 | 95 |
| $t = 13000$ | | | | |
| | * gaussian ($r_{L} = 1$), neigh | 0 | 0 | 0 | 100 |
| | * gaussian ($r_{L} = 1$) | 0 | 0 | 0 | 8 |
| | * gaussian ($r_{L} = 2$) | 0 | 0 | 0 | 0 |
| | * gaussian ($r_{L} = 2$), neigh | 0 | 0 | 0 | 100 |
| $t = 19500$ | | | | |
| | * gaussian ($r_{L} = 1$), neigh | 0 | 0 | 0 | 100 |
| | * gaussian ($r_{L} = 1$) | 0 | 0 | 0 | 7 |
| | * gaussian ($r_{L} = 2$) | 0 | 0 | 0 | 0 |
| | * gaussian ($r_{L} = 2$), neigh | 0 | 0 | 0 | 100 |
| $t = 26000$ | | | | |
| | * gaussian ($r_{L} = 1$), neigh | 0 | 0 | 0 | 100 |
| | * gaussian ($r_{L} = 1$) | 0 | 0 | 0 | 0 |
| | * gaussian ($r_{L} = 2$) | 0 | 0 | 0 | 0 |
| | * gaussian ($r_{L} = 2$), neigh | 0 | 0 | 0 | 100 |

Rank $r_{L} = 2$ Gaussian operators seem in fact to generate easier problems than rank $r_{L} = 1$ ones, at least judging from the given results. On the other hand, it becomes clear that exponentially decaying singular values pose significantly easier problems.
In numerical experiments, we have demonstrated that it can also practically suffice if only the number of Gaussian measurements exceeds the dimension of the lowest rank variety, the reference solution truth is contained in, by one. Further, we have shown that AIRLS-\(\mathcal{K}\) is only marginally less successful than its non-alternating version IRALS-\(\mathcal{K}\), while clearly superior towards ordinary, unregularized ALS. In moderately large cases, we could observe that 1.2 times the minimally necessary number of measurement in near all cases suffices to recover the solution. For large scale problems, it may yet show that a slower decline of \(\gamma\) could allow to further reduce the number of required measurements.

**Appendix A. (Remaining proof of Theorem 3.3)**

**Proof.** (of Theorem 3.3) Throughout the proof, we abbreviate \(W^\mathcal{K} := \{W^{(j)}\}_{J \in \mathcal{K}}\) as well as the iterates \(W^{(i)} := \{W^{(i,j)}\}_{J \in \mathcal{K}^{S_i}}\).

(i): Let \(\Delta^{(i)} = \gamma^{(i)} \sum_{J \in \mathcal{K}^{S_i}} n_J - \sum_{J \in \mathcal{K}^{S_i}} n_J\). Independent of \(S \subset \mathcal{K}\), we have

\[
\begin{align*}
\gamma^{(i)} & \geq \sum_{J \in \mathcal{K}^{S_i}} \left( \Delta^{(i)} + \|X^{(i+1)} - X^{(i)}\|_F^2 \right) \\
\end{align*}
\]

(ii): Since \((\text{cf. Section 2.1}) |\mathcal{K}| \gamma (\sum_{J \in \mathcal{K}} n_J)^{\alpha - 1} \|X\|_F^2 \leq \prod_{J \in \mathcal{K}} \prod_{i=1}^{n_J} (\sigma_i^{(J)}(X^2 + \gamma) \leq \exp(f_K^{(i)}(X))), it follows due to \((i)\) that \(|\mathcal{K}| \|X^{(i)}\|_F^2 \leq (\gamma^{(i)})^{1 - \sum_{J \in \mathcal{K}} n_J} \exp(f_K^{(i)}(X^{(i)}))).

As \(\gamma^{(i)}\) does not converge to zero, the sequence \(X^{(i)}\) remains bounded.

(iii): For \(S = S_i\) (and thus \(\Delta^{(i)} = 0\)), the steps \((d)\) to \((g)\) in \((i)\) provide that \(J^{K^{S_i}}_{(i)}(X^{(i+1)}, W^{(i)}) \geq f_{K^{S_i}}^{X^{(i+1)}}(X^{(i+1)}+1)\). With \(\mathcal{W}^{(i)}\) as defined in \((3.4)\), we then have

\[
\begin{align*}
J^{K^{S_i}}_{(i)}(X^{(i)}) - J^{K^{S_i}}_{(i+1)}(X^{(i+1)}) & \geq f_{K^{S_i}}^{X^{(i+1)}}(X^{(i)}, W^{(i)}) - f_{K^{S_i}}^{X^{(i+1)}}(X^{(i+1)}, W^{(i)}) \\
& = \langle X^{(i)}, \mathcal{W}^{(i)}(X^{(i)}) \rangle - \langle X^{(i+1)}, \mathcal{W}^{(i)}(X^{(i+1)}) \rangle \\
& = \langle X^{(i)} - X^{(i+1)}, \mathcal{W}^{(i)}(X^{(i)} + X^{(i+1)}) \rangle.
\end{align*}
\]

As \(\mathcal{W}^{(i)}(X^{(i+1)} + X^{(i+1)} \in \text{kernel} (\mathcal{L})\) (as provided by \((3.5)\)) we have

\[
\begin{align*}
\langle X^{(i)} - X^{(i+1)}, \mathcal{W}^{(i)}(X^{(i)} + X^{(i+1)}) \rangle & \geq \frac{1}{\lambda_{\min}^{(i)}} \|X^{(i)} - X^{(i+1)}\|_F^2 \lambda_{\min}^{(i)} \mathcal{W}^{(i)}
\end{align*}
\]

Since \(\mathcal{W}^{(i,j)} > 0, J \in \mathcal{K}^{S_i}\), the eigenvalue can be bounded via

\[
\lambda_{\min}^{(\mathcal{W}^{(i)})} = \lambda_{\min} \left( \sum_{J \in \mathcal{K}^{S_i}} \mathcal{W}^{(i,j)} \right) \geq \sum_{J \in \mathcal{K}^{S_i}} \lambda_{\min}(\mathcal{W}^{(i,j)})
\]

\[
= \sum_{J \in \mathcal{K}^{S_i}} \lambda_{\min}(\sigma^2_{\gamma}(X^{(i)})^2 + \gamma)^{-1}
\]

\[
= \sum_{J \in \mathcal{K}^{S_i}} \lambda_{\min}(\sigma^2_{\gamma}(X^{(i)})^2 + \gamma)^{-1} \geq |\mathcal{K}| \|X^{(i)}\|_F^2 + \gamma)^{-1}
\]

Therefore, as \(\|X^{(i)}\|_F^2\) remains bounded due to \(\gamma^* > 0\) and \((ii)\), there exists \(c > 0\) such that \(\|\|(X^{(i)} - X^{(i+1)})\|_F^2 \lambda_{\min}^{(\mathcal{W}^{(i)})} \geq c\|\|(X^{(i)} - X^{(i+1)})\|_F^2\). Summing over all
i = 1, . . . , N, we obtain
\[ \sum_{S \in K} c_i N \| (X^{(i)} - X^{(i+1)}) \|_F^2 \leq \sum_{S \in K} f^{S}_{\gamma_i}(X^{(i)}) - f^{S}_{\gamma_{i+1}}(X^{(i+1)}) \]
\[ \leq \sum_{S \in K} f^{S}_{\gamma_0}(X^{(i)}) - f^{S}_{\gamma_0}(X^{(i+1)}) = \sum_{S \in K} f^{S}_{\gamma_0}(X^{(1)}) - f^{S}_{\gamma_0}(X^{(N+1)}). \]

As for each \( S \subset K \), \( f^{S}_{\gamma_0}(X^{(i)}) \) remains bounded, this implies \( \| (X^{(i)} - X^{(i+1)}) \|_F^2 \rightarrow 0 \) for \( i \rightarrow \infty \).

(iii/2): This part is largely independent of choices of \( S \subset K \) since the stationary points of all \( f^{\delta}_{S} \) are equal (cf. Section 3.2). Let \( X^{(i)} \) be a convergent subsequence of \( X^{(i)} \) with limit point \( X^{*} \). In light of Corollary 3.1, it suffices to show that \( X^{*} = X^{(0)}_{W^{*}} \) for \( W^{(i)} = W^{(J)}_{\gamma^{(i)}}, X^{(i)} \), \( J \in K^{S} \) for one \( S \subset K \). Due to (iii/1) so far, we have \( \lim_{i \rightarrow \infty} X^{(i+1)} = X^{*} \). As \( W^{(J)}_{\gamma}, X^{(i)} \), \( J \in [d] \), depend continuously on \( X \) and \( \gamma > 0 \), it follows that
\[ W^{(i)} = W^{(J)}_{\gamma^{(i)}}, X^{(i)} \rightarrow_{i \rightarrow \infty} W^{(J)}_{\gamma^{*}, X^{*}} =: W^{*}. \]

Let now \( S \) be one of the sets that appear infinitely often in \( \{S_i\}_{i \in \mathbb{N}} \) with respect to a subsequence \( \{i_k\}_{k \in \mathbb{N}}, S_{i_k} = S, k \in \mathbb{N} \). Then as \( X^{(i)}_{W^{*}} \) depends continuously on \( W^{(J)} \), \( J \in K^{S} \), the last remaining step is shown by
\[ X^{*} \leftarrow_{k \rightarrow \infty} X^{(i_{k+1})} = X^{(i_{k+1})}_{W^{*}_{i_k}} \rightarrow_{k \rightarrow \infty} X^{(i_k)}_{W^{*}}. \]

(iv): This part is word for word the same as in [26].

\[ \square \]

Acknowledgments

The author would like to thank Maren Klever and Lars Grasedyck for fruitful discussions, as well as Paul Breiding and Nick Vannieuwenhoven for conversations on generic recoverability within varities.

References

[1] J. Ballani, L. Grasedyck, and M. Kluge, Black box approximation of tensors in hierarchical Tucker format, Linear Algebra and its Applications, 438 (2013), pp. 639 – 657.
[2] C. Bayer, M. Eigel, L. Sallandt, and P. Trunschke, Pricing high-dimensional Bermudan options with hierarchical tensor formats, 2021.
[3] P. Breiding, T. O. Çelik, T. Duff, A. Heath, A. Maraj, A.-L. Sattelberger, L. Venturino, and O. Yüüük, Nonlinear algebra and applications, 2021.
[4] P. Breiding, F. Gesmundo, M. Michalek, and N. Vannieuwenhoven, Algebraic compressed sensing (in preparation), 2021.
[5] E. J. Candés, M. B. Wakin, and S. P. Boyd, Enhancing sparsity by reweighted l1 minimization, Journal of Fourier Analysis and Applications, 14 (2008), pp. 877–905.
[6] L. Chiantini, G. Ottaviani, and N. Vannieuwenhoven, An algorithm for generic and low-rank specific identifiability of complex tensors, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 1265–1287.
[7] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, Communications on Pure and Applied Mathematics, 63 (2010), pp. 1–38.
[8] L. De Lathauwer, A survey of tensor methods, in 2009 IEEE International Symposium on Circuits and Systems (ISCAS), May 2009, pp. 2773–2776.
[9] L. De Lathauwer, B. De Moor, and J. Vandewalle, A multilinear singular value decomposition, SIAM Journal on Matrix Analysis and Applications, 21 (2000), pp. 1253–1278.
[10] A. Falcó, W. Hackbusch, and A. Nouy, Tree-based tensor formats, SeMA Journal, (2018).
[11] M. Fornasier, H. Rauhut, and R. Ward, Low-rank matrix recovery via iteratively reweighted least squares minimization, SIAM Journal on Optimization, 21 (2011), pp. 1614–1640.
[12] S. Gandy, B. Recht, and I. Yamada, Tensor completion and low-rank tensor recovery via convex optimization, Inverse Problems, 27 (2011), p. 025010.
[13] A. Goessmann, M. Gotte, I. Roth, R. Sweke, G. Kutyniok, and J. Eisert, Tensor network approaches for learning non-linear dynamical laws, 2020.
[14] M. Gotte, R. Schneider, and P. Trunschke, A block-sparse tensor train format for sample-efficient high-dimensional polynomial regression, 2021.
[15] L. Grasedyck, Hierarchical singular value decomposition of tensors, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 2029–2054.
[16] L. Grasedyck and S. Krämer, Stable als approximation in the tt-format for rank-adaptive tensor completion, Numerische Mathematik, (2019).
[17] L. Grasedyck, D. Kressner, and C. Tobler, A literature survey of low-rank tensor approximation techniques, GAMM-Mitteilungen, 36 (2013), pp. 53–78.
[18] E. Grelier, A. Nouy, and M. Chevreuil, Learning with tree-based tensor formats, 2019.
[19] C. Haberstich, A. Nouy, and G. Perrin, Active learning of tree tensor networks using optimal least-squares, 2021.
[20] S. Holtz, T. Rohwedder, and R. Schneider, The alternating linear scheme for tensor optimization in the tensor train format, SIAM Journal on Scientific Computing, 34 (2012), pp. A683–A713.
[21] S. Holtz, T. Rohwedder, and R. Schneider, On manifolds of tensors of fixed tt-rank, Numerische Mathematik, 120 (2012), pp. 701–731.
[22] Y. Kapushev, I. Oseledets, and E. Burnaev, Tensor completion via gaussian process-based initialization, SIAM Journal on Scientific Computing, 42 (2020), pp. A3812–A3824.
[23] T. G. Kolda and B. W. Bader, Tensor decompositions and applications, SIAM Review, 51 (2009), pp. 455–500.
[24] S. Krämer, A geometric description of feasible singular values in the tensor train format, SIAM Journal on Matrix Analysis and Applications, 40 (2019), pp. 1153–1178.
[25] S. Krämer, Tree tensor networks, associated singular values and high-dimensional approximation, dissertation, RWTH Aachen University, Aachen, 2020. Veröffentlicht auf dem Publikationsserver der RWTH Aachen University; Dissertation, RWTH Aachen University, 2020.
[26] S. Krämer, Asymptotic log-det rank minimization via (alternating) iteratively reweighted least squares, 2021.
[27] D. Kressner, M. Steinlechner, and B. Vandereycken, Low-rank tensor completion by nonconvex optimization, BIT Numerical Mathematics, 54 (2014), pp. 447–468.
[28] Y. Liu and F. Shang, An efficient matrix factorization method for tensor completion, IEEE Signal Processing Letters, 20 (2013), pp. 307–310.
[29] K. Mohan and M. Fazel, Iterative reweighted algorithms for matrix rank minimization, Journal of Machine Learning Research, 13 (2012), pp. 3441–3473.
[30] A. Nouy, Low-Rank Tensor Methods for Model Order Reduction, Springer International Publishing, Cham, 2017, pp. 857–882.
[31] I. Oseledets and E. Tyrtyshnikov, Tt-cross approximation for multidimensional arrays, Linear Algebra and its Applications, 432 (2010), pp. 70 – 88.
[32] I. V. Oseledets, Tensor-train decomposition, SIAM Journal on Scientific Computing, 33 (2011), pp. 2295–2317.
[33] Y. Qi, P. Comon, and L.-H. Lim, Semialgebraic geometry of nonnegative tensor rank, SIAM Journal on Matrix Analysis and Applications, 37 (2016), pp. 1556–1580.
[34] H. Rauhut, R. Schneider, and Z. Stojanac, Tensor Completion in Hierarchical Tensor Representations, Springer International Publishing, Cham, 2015, pp. 419–450.
[35] M. Signoretto, Q. Tran-Dinh, L. De Lathauwer, and J. A. K. Suykens, Learning with tensors: a framework based on convex optimization and spectral regularization, Machine Learning, 94 (2014), pp. 303–351.
[36] C. D. Silva and F. J. Herrmann, Optimization on the hierarchical Tucker manifold – applications to tensor completion, Linear Algebra and its Applications, 481 (2015), pp. 131 – 173.
[37] M. Sørensen and L. De Lathauwer, Fiber sampling approach to canonical polyadic decomposition and application to tensor completion, SIAM Journal on Matrix Analysis and Applications, 40 (2019), pp. 888–917.
[38] M. Sørensen, N. D. Sidiropoulos, and L. De Lathauwer, Canonical polyadic decomposition of a tensor that has missing fibers: A monomial factorization approach, in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 7490–7494.
[39] M. Steinlechner, Riemannian optimization for high-dimensional tensor completion, SIAM Journal on Scientific Computing, 38 (2016), pp. S461–S484.
[40] L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311.

[41] A. Uschmajew and B. Vandereycken, *The geometry of algorithms using hierarchical tensors*, Linear Algebra and its Applications, 439 (2013), pp. 133–166.

[42] G. Vidal, *Efficient classical simulation of slightly entangled quantum computations*, Phys. Rev. Lett., 91 (2003), p. 147902.
SUPPLEMENTARY MATERIALS:
SM1. Alternating ASRM (further experiment)

Experiment SM1. For \(d = 4, \pi = 5, \tau_{rs} = 3\) and \(\ell \in \{69, 83, 111, 138\}\), we consider the ASRM-\(\mathcal{K}_{bbHT}\) problem based on samples for reference solutions given via \(bbHT\) representations. We use the following four solution methods:

(a) full, image based (as already considered in Experiment 7.2)
(b) full, relaxed
(c) alternating, based on fixed ranks \(r^{(J)} = 5, J \in \mathcal{K}_{bbHT}\)
(d) alternating, with adaptive ranks \(r^{(J)} \in [5], J \in \mathcal{K}_{bbHT}\) (cf. Section SM6.3)

Each constellation is repeated 100 times, for \(k_{\text{max}} = 8\). The results are covered in Table SM1 and Figs. SM5 and SM6.

### Table SM1. (A)IRLS-\(0\mathcal{K}_{bbHT}\) (sampling, \(d = 4, \pi = 5, \tau_{rs} = 3\)) – table as specified in Section 7.5 for Experiment SM1 (see Fig. SM6 for more details)

| instance \(\ell\) | \(\text{ASRM-}\mathcal{K}_{bbHT}\) recovery | \(Q_x \in [0, 0.98]\) | \(Q_x \in (0.98, 1.005]\) | \(Q_x \in [1.005, \infty]\) | \(Q_x = \infty\) |
|-------------------|--------------------------------------|-----------------|-----------------|-----------------|-----------------|
| \(\ell = 69\)     | (a) full, image                       | 0               | 0               | 0               | 100             |
|                   | (b) full, relaxed                     | 0               | 0               | 0               | 100             |
|                   | (c) alternating, rank fixed           | 0               | 0               | 0               | 100             |
|                   | (d) alternating, rank adapted         | 0               | 0               | 0               | 100             |
| \(\ell = 83\)     | (a) full, image                       | 0               | 0               | 33              | 67              |
|                   | (b) full, relaxed                     | 0               | 0               | 33              | 67              |
|                   | (c) alternating, rank fixed           | 0               | 0               | 15              | 85              |
|                   | (d) alternating, rank adapted         | 0               | 0               | 14              | 86              |
| \(\ell = 111\)    | (a) full, image                       | 0               | 0               | 94              | 6               |
|                   | (b) full, relaxed                     | 0               | 0               | 94              | 6               |
|                   | (c) alternating, rank fixed           | 0               | 0               | 96              | 4               |
|                   | (d) alternating, rank adapted         | 0               | 0               | 95              | 5               |
| \(\ell = 138\)    | (a) full, image                       | 0               | 0               | 100             | 0               |
|                   | (b) full, relaxed                     | 0               | 0               | 100             | 0               |
|                   | (c) alternating, rank fixed           | 0               | 0               | 100             | 0               |
|                   | (d) alternating, rank adapted         | 0               | 0               | 100             | 0               |

There does not seem to be a relevant difference between full image based or relaxed optimization. Further, only for \(\ell = 83\) alternating optimization performs slightly worse for. The explicit adaption of the rank in turn likewise yields no notable difference. The quality of approximation is thus seemingly only reduced (and only slightly so) through the change to an alternating optimization. However, this effect might go stronger with increased dimensions \(d\).
SM2. Visualization of numerical results

Each of the following even and odd numbered pair of pages contains two related visualizations of the results of one of experiments SM1 and 7.2 to 7.4 as summarized in Table SM2. These additional visualizations are constructed as described further below.

| experiment | γ-sensitivity | ASRM/recovery | table |
|------------|---------------|---------------|-------|
| Experiment 7.1 | Fig. SM1 | Fig. SM2 | Table 1 |
| Experiment 7.2 | Fig. SM3 | Fig. SM4 | Table 2 |
| Experiment SM1 | Fig. SM5 | Fig. SM6 | Table SM1 |
| Experiment 7.3 | (ν = 1.002−1) | Fig. SM7 | Table 3 |
| Experiment 7.4 | Fig. SM8 | Fig. SM9 | Table 4 |
| Experiment 7.4  (s_{(expfac)} = 1/3) | Fig. SM10 | Fig. SM11 | Table 5 |

Table SM2. overview over experiments, related figures and tables

γ-decline sensitivity. To each single trial that did not yield a failure, we assign the one index $k$ for which the parameter $\nu = \nu_k$ first led to a successful or improving run as described in Section 7.4. The frequencies of these indices as well as fails are then plotted as bars, where improvements are plotted below the x-axis.

ASRM/recovery figures. We display the following points as button plot (as defined below). Given the $i$-th result $X^{(alg)}$ as well as reference solution $X^{(rs)}$, the x-value of the $i$-th point is given by the bounded quotient

$$x_i = \max(0.9, \min(Q_e(X^{(alg)}, X^{(rs)}), 1.05)),$$

Each y-value is given by

$$y_i = \min(\|X^{(alg)} - X^{(rs)}\|_F/\|X^{(rs)}\|_F, 1),$$

Note that the algorithm stops automatically if that value falls below $10^{-6}$.

button plot. With a button plot (with logarithmic scale in y), we refer to a two dimensional, clustered scatter plot. Therein, any circular markers with centers $(x_i, y_i)$ and areas $s_i$, $i = 1, \ldots, k$, that would (visually) overlap, are recursively combined to each one larger circle $(\hat{x}, \hat{y})$ with area $\hat{s}$ according to the appropriately weighted means

$$\hat{x} = \frac{\sum_{i=1}^{k} s_i x_i}{\hat{s}}, \quad \hat{y} = \prod_{i=1}^{k} y_i^{s_i/\hat{s}}, \quad \hat{s} = \sum_{i=1}^{k} s_i.$$

The centers of all resulting circles are indicated as crosses. Thus, if only one circle remains, then the position of that cross is given by the arithmetic mean of all initial x-coordinates and the geometric mean of all initial y-coordinates. If no disks are combined, then their centers are the initial coordinates and their areas are all equal.
Figure SM1. Results for Experiment 7.1 as described in Section SM2.

tensor recovery (full, Gaussian, image method, $d = 4, r = 5, r_{\text{as}} = 3$)
tensor recovery (full, Gaussian, image method, $d = 4$, $n = 5$, $r_{sa} = 3$)

$\Delta \ell = -1$

HT, $\ell = 68$

$\Delta \ell = 0$

HT, $\ell = 69$

$\Delta \ell = 1$

HT, $\ell = 70$

Figure SM2. Results for Experiment 7.1 as described in Section SM2.
IRLS-0K (full, image method, $d = 4$, $\pi = 5$, $\tau_n = 3$)

$c_{\text{mf}} = 1.2$

Figure SM3. Results for Experiment 7.2 as described in Section SM2.
IRLS-0K (full image method, $d = 4$, $\pi = 5$, $\tau_m = 3$)

gaussian, $\ell = 83$

gaussian, $\ell = 62$

sampling, $\ell = 83$

sampling, $\ell = 62$

gaussian, $\ell = 111$

gaussian, $\ell = 82$

sampling, $\ell = 111$

sampling, $\ell = 82$

gaussian, $\ell = 138$

gaussian, $\ell = 102$

sampling, $\ell = 138$

sampling, $\ell = 102$

Figure SM4. Results for Experiment 7.2 as described in Section SM2.
Figure SM5. Results for Experiment SM1 as described in Section SM2.
(A) HLS-0KART (sampling, \(d = 4, \pi = 5, r_n = 3\))

\[ \ell = 69 \]

\[ \ell = 83 \]

\[ \ell = 111 \]

\[ \ell = 138 \]

Figure SM6. Results for Experiment SM1 as described in Section SM2.
ASYMPTOTIC LOG-DET SUM-OF-RANKS MINIMIZATION VIA TENSOR (A)IRLS SM9

(A)IRLS-0K Tucker \( (Tucker, d = 4, \sigma = 5, \tau_m = 3) \)

\( t = 126 \)

\begin{align*}
\text{gaussian (72 - 1): (a) full} & \quad \text{(d) alt adapt} \quad \text{(e) plain ALS} \\
\text{gaussian (72 - 1): (a) full} & \quad \text{(d) alt adapt} \quad \text{(e) plain ALS} \\
\text{gaussian (72 - 1): (a) full} & \quad \text{(d) alt adapt} \quad \text{(e) plain ALS}
\end{align*}

Figure SM7. Results for Experiment 7.3 as described in Section SM2.
(A) IRLS-0K\textsubscript{bbHT} (alternating with heuristics, \(d = 8, \pi = 20, \pi_{\text{rs}} = 5\))

\(\ell = 1\,3000\)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig6/callback5_el13000.png}
\caption{Results for Experiment 7.4 as described in Section SM2.}
\end{figure}

\(\ell = 1\,9500\)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig6/callback5_el19500.png}
\caption{Results for Experiment 7.4 as described in Section SM2.}
\end{figure}

\(\ell = 2\,6000\)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig6/callback5_el26000.png}
\caption{Results for Experiment 7.4 as described in Section SM2.}
\end{figure}
ASYMPTOTIC LOG-DET SUM-OF-RANKS MINIMIZATION VIA TENSOR (A)IRLS

SM11

(A)IRLS-0K_{syst} (alternating, with heuristics, d = 8, π = 20, β = 5)

\[ \ell = 13000 \]

\[ \text{gaussian } (r_L = 1) \]
\[ \text{gaussian } (r_L = 1), \text{ neigh} \]
\[ \text{gaussian } (r_L = 2) \]
\[ \text{gaussian } (r_L = 2), \text{ neigh} \]

\[ \ell = 19500 \]

\[ \text{gaussian } (r_L = 1) \]
\[ \text{gaussian } (r_L = 1), \text{ neigh} \]
\[ \text{gaussian } (r_L = 2) \]
\[ \text{gaussian } (r_L = 2), \text{ neigh} \]

\[ \ell = 26000 \]

\[ \text{gaussian } (r_L = 1) \]
\[ \text{gaussian } (r_L = 1), \text{ neigh} \]
\[ \text{gaussian } (r_L = 2) \]
\[ \text{gaussian } (r_L = 2), \text{ neigh} \]

Figure SM9. Results for Experiment 7.4 as described in Section SM2.
(A) IRLS-0KDAHT (alternating, with heuristics, exp. dec. \(s = 1/3\), \(d = 8\), \(\sigma = 20\), \(\tau_0 = 5\))

\[\ell = 6500\]

\[\ell = 13000\]

\[\ell = 19500\]

\[\ell = 26000\]

Figure SM10. Results for Experiment 7.4 as described in Section SM2.
Figure SM11. Results for Experiment 7.4 as described in Section SM2.
SM4. Proof of Theorem 6.3

Following is the proof of Theorem 6.3 in minimally deviating notation. For the more elegant version, see Section SM5.2.

Proof. Firstly, we consider a $|E_Y| + |E_c|$ dimensional tensor representation $G^{(J)} \in \mathbb{R}^{X_{c} \times X_{y} \times |J|} \otimes \mathbb{D}_{m_e}$ of $N_{\beta \epsilon} : \mathbb{D}_{m_e} \rightarrow \mathbb{R}^{n_1 \times \ldots \times n_d}$ and its adjoint. We can thus write

$$A^{(J)}_{\alpha', \beta'} = \sum_{\alpha, \beta} G^{(J)}_{\alpha, \beta} W^{(J)}_{\alpha, \beta} \delta^{(J)}_{\alpha, \beta'}$$

where the path evaluation is given by

$$G^{(J)}_{\alpha, \epsilon} = \sum_{\beta, \epsilon} \prod_{\alpha, \epsilon} Y^{(J)}_{\alpha, \beta}$$

The representation $G^{(J)}$ can further be decomposed into a set of orthonormal matrices $Y^{(J)} \in \mathbb{R}^{[n_c] \times [p^{(J)}]_c}$, $e \in E_Y$, and the tensor $P^{(J)} \in \mathbb{R}^{X_{c} \times X_{y} \times \mathbb{D}_{m_c}[p^{(J)}_c]}$ obtained via a contraction along the path $p$,

$$\frac{\partial}{\partial \epsilon} P^{(J)}_{\alpha, \beta} = \sum_{\beta', \epsilon} \prod_{\alpha, \epsilon} Y^{(J)}_{\alpha, \beta}$$

As Lemma 6.2 provides, we may replace $W^{(J)} = W^{(J)}_{\gamma, \tilde{N}, \epsilon}$. The matrices $Y^{(J)}$, $\gamma \in \mathbb{K}^{E}$, then cancel out due to orthonormality and we obtain (6.6) for

$$M^{(J)}_{\beta', \beta} = \sum_{\beta', \beta : e \in E_Y \setminus E_c} \prod_{\alpha, \epsilon} \delta^{(J)}_{\beta', \beta}$$

As the term $H^{(J)}$ can similarly be simplified, we have

$$H^{(J)}_{\beta', \beta} = \sum_{\beta', \beta : e \in E_Y \setminus \{\tilde{e}\}} \prod_{\alpha, \epsilon} \delta^{(J)}_{\beta', \beta}$$

where

$$P^{(J)}_{\alpha, \beta} = \sum_{\beta, \epsilon} (N_{\epsilon})_{\alpha, \beta}$$

By expanding and reordering the contractions within the path evaluations, we then arrive at (6.7) and (6.8). \qed
SM5. Tensor nodes

As indicated in Section 5.1, we in the following dismiss the indices in tensor contractions. What is here introduced as notation, is a simplified version of the formal arithmetic established in [25].

SM5.1. Self-emergent contractions. Though it is clear by Section 5, which tensors are assigned which labels, we here repeat this formal step. Such is indicated by writing \( X = X(\{\alpha_\mu\}_{\mu \in [d]} \in \mathcal{F}_{\{\alpha_\mu\}_{\mu \in [d]}} \) for any full tensor, or in case of its representation network \( X = \tau_r(N) \), by

\[
N = \{N_v\}_{v \in V}, \quad N_v = N_v(\{\gamma\}_{\gamma \in m_v}) \in \mathcal{H}_{m_v}.
\]

Avoiding the redundant notation such as in the expression (5.5), we simply write

\[
X = \bigotimes_{v \in V} N_v.
\]

The same symbol is used for any other contraction, such as in (5.6), translating to

\[
\tau_r(\{N_s\}_{s \in S})_{(\alpha_s)_{\in S}} \equiv \bigotimes_{v \in S} N_s.
\]

For any label \( \gamma \), we denote the priorily used Kronecker deltas as formal objects \( \Delta_{\gamma, \gamma} \in \mathcal{H}_{\{\gamma, \gamma\}} \), or equivalently so for more than two labels. Instead of explicitly denoting primed labels, we instead define

\[
(N_v)_{\gamma \rightarrow \gamma'} := \Delta_{\gamma, \gamma'} \otimes N_v, \quad \gamma \in m_v.
\]

As shorthand notation, we further define \( N'_v := N'_v(\alpha_v, \{\beta^e\}_{e \in E_v}) \) as

\[
N'_v := (N_v)_{\{\beta^e\}_{e \in E_v} \rightarrow \{\beta^e\}_{e \in E_v}} := \bigotimes_{e \in E_v} \Delta_{\beta^e, \beta^e} \otimes N_v.
\]

For other tensors, the operator (\( \cdot' \)) likewise denotes a priming of all labels \( \{\beta^e\}_{e \in E} \) assigned to such. The special case of an element-wise multiplication as in (6.4) is flagged via a superindex

\[
L_v \bigotimes L_v := \Delta_{\nu', \cdot', \cdot} \otimes (\Delta_{\cdot', \cdot} \bigotimes L_v) \otimes (\Delta_{\nu', \cdot} \bigotimes L_v).
\]

We may thus equivalently write (6.4) as

\[
\mathcal{L}(X) = L \bigotimes X \in \mathcal{H}_\xi, \quad L = \rho_{KL}(\{L_v\}_{v \in V}) = \bigotimes_{v \in V} L_v \in \mathcal{H}_{\xi \cup \{m_v\}_{v \in [d]}},
\]

The expression (6.5) for instance takes the shorter shape \( \mathcal{L}(X) = \bigotimes_v (L_v \bigotimes N_v) \).

SM5.2. Alternative Theorem 6.3 and Corollary 6.4. While we may write

\[
A^{(j)} = A^{(j)}(\alpha'_v; \{\beta^e\}_{e \in E_v}, \alpha_c; \{\beta^c\}_{e \in E_c}),
\]

the identities in Theorem 6.3 become

(SM1) \[
A^{(j)} = \Delta_{\alpha'_v, \cdot} \bigotimes \left( \bigotimes_{e \in E_v \setminus \{e_1\}} \Delta_{\beta^e, \beta^e} \right) \bigotimes M^{(j)},
\]

for \( M^{(j)} = M^{(j)}(\beta^{e_1}, \beta^{c_1}) \) with

(SM2) \[
M^{(j)} = \left( \bigotimes_{e \in E_v \setminus \{e_1\}} \Delta_{\beta^e, \beta^e} \bigotimes \left( \bigotimes_{v \in P} N'_v \bigotimes N_v \right) \right) \bigotimes (H^{(j)} + \gamma I)^{-1},
\]

as well as \( H^{(j)} = H^{(j)}(\beta^{e_1}, \beta^{c_1}) \) given by

(SM3) \[
H^{(j)} = \left( \bigotimes_{e \in E_v \setminus \{e\}} \Delta_{\beta^e, \beta^e} \right) \bigotimes \left( \bigotimes_{v \in \{v\} \cup P} N'_v \bigotimes N_v \right).
\]

The identities appearing in the proof of Theorem 6.3, in turn, become

\[
A^{(j)} = G^{(j)} \bigotimes W^{(j)} \bigotimes G^{(j)}, \quad G^{(j)} = Y(J_v) \bigotimes P^{(j)}, \quad P^{(j)} = \bigotimes_{v \in P} N_v.
\]
These tensors thereby have labels $G^{(j)} = G^{(j)}(\{\alpha_c\}_{c \in E_Y}, \alpha_c, \{\beta^e\}_{e \in E_v})$, as well as $\Phi^{(j)} = \Phi^{(j)}(\alpha_{c'}, \beta^e), e \in E_Y$, and $P^{(j)} = P^{(j)}(\{\alpha_c\}_{c \in V}, \{\beta^e\}_{e \in \partial E_y})$. Further,
\[
M^{(j)} = P^{(j)}_{\beta^1 \rightarrow \beta^1, \beta^2 \rightarrow \beta^2} \otimes H^{(j)} \otimes P^{(j)}
\]
and similarly
\[
H^{(j)} = P^{(+e, j)}_{\beta^2 \rightarrow \beta^2} \otimes P^{(+e, j)} \cdot P^{(+e, j)} = N_{c} \otimes P^{(j)}
\]
The identity in Corollary 6.4 on the other hand is simply $H^{(j)} = (N_{c})_{\beta^2 \rightarrow \beta^2} \otimes N_{c}$.

SM5.3. **Proofs of Propositions 6.5 to 6.7.** The recursion stated in Proposition 6.5 is
\[
F_c = L_c \otimes_{v \in \text{neigh}(c)} S^{(j(c,v))},
\]
for $c \in V$ and
\[
S^{(j)} = (L_v \otimes N_v) \otimes_{b \in \text{desc}_c(v)} S^{(j(v,b))},
\]
with $\hat{e} = \{\text{pred}_c(v), v\}$ for $v \in V \setminus \{c\}$.

**Proof.** The recursion implies that $S^{(j)} = \otimes_{v \in \text{branch}_c(v)} (L_h \otimes N_h)$. Thereby,
\[
F_c := L_c \otimes_{v \in V \setminus \{c\}} (L_v \otimes N_v) = L_c \otimes_{v \in \text{neigh}(c)} \otimes_{h \in \text{branch}_c(v)} (L_h \otimes N_h)
\]
provides the to be shown, first identity. $\square$

The recursion in Proposition 6.6 is
\[
\sum_{J \in \mathcal{K}} A^{(j)} = \Delta_{\alpha'_e, \alpha_e} \otimes \sum_{v \in \text{neigh}(c)} \left( e \in E_v \otimes \{j, e\} \right) \Delta_{\beta^{e'}, \beta^e} \otimes B^{(j(v,c))},
\]
where for $\hat{e} = \{p_{-1}, v\}, p_{-1} = \text{pred}_c(v), v \in V \setminus \{c\}$, it is
\[
\hat{B}^{(j(v,c))} = (N_v)_{\beta^{e'} \rightarrow \beta^{e'}, \beta(v,c) \rightarrow \beta(v,h)} \otimes B^{(j(v,c))} \otimes N_v,
\]
for $b \in \text{desc}_c(v)$.

**Proof.** By definition,
\[
\sum_{J \in \mathcal{K}} A^{(j)} = \sum_{e \in V \setminus \{c\}} \Delta_{\alpha'_e, \alpha_e} \otimes \left( e \in E_v \setminus \{c, p_{1}\} \right) \Delta_{\beta^{e'}, \beta^e} \otimes M^{(j(p_{1}, c))}
\]
\[
= \Delta_{\alpha'_e, \alpha_e} \otimes \sum_{v \in \text{neigh}(c)} \left( e \in E_v \setminus \{c, p_{1}\} \right) \Delta_{\beta^{e'}, \beta^e} \otimes \sum_{b \in \text{branch}_c(v)} M^{(j(b, c))}.
\]
Let each be $p = (c, h)$, with $e_1 = \{c, p_{1}\}, \hat{e} = \{p_{-1}, h\}$, for $h \in V \setminus \{c\}$. We show that
\[
\sum_{b \in \text{branch}_c(h)} M^{(j(p_{1}, c))} = \left( e \in \partial E_v \setminus \{e_1, \hat{e}\} \right) \Delta_{\beta^{e_1}, \beta^e} \otimes \left( \otimes_{v \in \hat{e}} N_v \otimes N_{v} \right) \otimes B^{(j)}
\]
by induction over the cardinality of $\text{branch}_c(h)$. The induction start for a cardinality of 1 is then given by (SM2) and the definition of $B^{(j)}$. In turn, given the tree
structure of \( G \) and the induction hypothesis, it follows that
\[
\sum_{b \in \text{branch}_c(h)} M^{(J_{\text{pred}_c(b),a})} = M^{(J_c)} + \sum_{b \in \text{desc}_c(h)} \sum_{w \in \text{branch}_h(b)} M^{(J_{\text{pred}_w(c),w})}
\]
\[
= \left( \sum_{e \in \partial E_{v} \setminus \{e_c, v\}} \Delta_{\ell^{c,v}, b_e} \right) \boxtimes \left( \sum_{v \in p} N_v' \boxtimes N_v \right) \otimes (H^{(J_c)} + \gamma I)^{-1}
\]
\[
+ \sum_{b \in \text{desc}_c(h)} \left( \sum_{e \in E_{v} \setminus \{e_c, h\}} \Delta_{\ell^{c,v}, b_e} \right) \boxtimes \left( \sum_{v \in e_c, b} N_v' \boxtimes N_v \right) \otimes B^{(J_{h,b})} \boxtimes N_h.
\]

The last to be shown step follows as by definition
\[
B^{(J_c)} = (H^{(J_c)} + \gamma I)^{-1} + \sum_{b \in \text{desc}_c(h)} (N_h)_{\beta^{c,v} \rightarrow \beta^{c,v}, \beta(h,b)} \rightarrow \beta(h,a)} \otimes B^{(J_{h,b})} \boxtimes N_h.
\]

For \( e \in E_{v} \setminus \{e_c, v\} \), \( p_{-1} = \text{pred}_c(v) \), \( v \in V \setminus \{e \} \), is
\[
H^{(J_{v,v})} = (N_v)_{\beta^{v,v} \rightarrow \beta^{v,v}, \beta^{v,v} \rightarrow \beta^{v,v}} \otimes H^{(J_v)} \boxtimes N_v,
\]

for \( v \in \text{desc}_c(v) \).

\[\]

Proof. Using the identity (SM3) on both sides, we obtain
\[
(N_v)_{\beta^{v,v} \rightarrow \beta^{v,v}, \beta^{v,v} \rightarrow \beta^{v,v}} \otimes H^{(J_v)} \boxtimes N_v
\]
\[
= (N_v)_{\beta^{v,v} \rightarrow \beta^{v,v}, \beta^{v,v} \rightarrow \beta^{v,v}} \otimes \left( \sum_{e \in \partial E_{v} \setminus \{e\}} \Delta_{\ell^{v,v}, b_e} \right) \boxtimes \left( \sum_{w \in \{v\}} N_w' \boxtimes N_w \right) \boxtimes N_v
\]
\[
= \left( \sum_{e \in E_{v} \setminus \{e, v\}} \Delta_{\ell^{v,v}, b_e} \right) \boxtimes N_v' \otimes \left( \sum_{e \in E_{v} \setminus \{e\}} \Delta_{\ell^{v,v}, b_e} \right) \boxtimes \left( \sum_{w \in \{v\}} N_w' \boxtimes N_w \right) \boxtimes N_v
\]
\[
= \left( \sum_{e \in E_{v} \setminus \{e, v\}} \Delta_{\ell^{v,v}, b_e} \right) \boxtimes \left( \sum_{w \in \{v\}} N_w' \boxtimes N_w \right) = H^{(J_{v,v})},
\]

which was to be shown. \(\square\)
**SM5.4. Detailed AIRLS-0K algorithm.** Algorithm 3 summarizes the AIRLS-0K method as covered in Section 6. In our experiments, we have chosen the therein appearing constant as $c_E = \frac{1}{4} |K|^{-1} \|L\|_F^2 / m_d$. The heuristics laid out in Section SM6 are marked as possibly applicable statements.

\begin{algorithm}
\caption{Detailed AIRLS-0K method}
1: derive tree $G = (V, E)$ from $K$
2: set $N = \{N \}_v \in V \in D_{\tau}$, $\gamma(0) > 0$, $c_0 \in V$
3: possibly introduce validation set (cf. Section SM6.1)

// let $J_e \in K_{S_{0o}}$, $e \in E$
// orthonormalize $\{N_v \}_v \in V$ with respect to $c_0$ (cf. Theorem 5.11) and initialize the branch evaluations $\{S^{(J_e)} \}_e \in E$ (cf. Proposition 6.5)

4: $\tilde{V} := \text{leaves}(c_0)$
5: while $\tilde{V} \neq \{c_0\}$ do
6: \hspace{1em} for $v \in \tilde{V} \setminus \{c_0\}$ do
7: \hspace{2em} $p_{-1} := \text{pred}_{c_0}(v)$, $\hat{e} := \{p_{-1}, v\}$
8: \hspace{2em} $QR := N^{|v\setminus \{\hat{e}\}}_v, N^{|v\setminus \{\hat{e}\}}_v := Q_N \gamma_p^{\hat{e}} := R_N^{\hat{e}}$
9: \hspace{2em} $S^{(J_e)} := (L_0 \boxtimes N_v) \boxtimes_{\hat{e} \in \text{pred}_{c_0}(v)} S^{(J_e; \hat{e}, \hat{e})}$
10: \hspace{2em} end for
11: $\tilde{V} := \bigcup_{v \in \tilde{V}} \text{pred}_{c_0}(v)$
12: end while

// implicitly declare iterate $X^{(0)} = \tau_r(\{N_v \}_v \in V)$
13: for $i = 1, 2, \ldots$ do
14: \hspace{1em} // calculate $\{H^{(i-1)} \}_e \in K_{S_{i-1}}$ (cf. Proposition 6.7 and Corollary 6.4) or possibly limit set via maximal distance of $\tilde{v} \in \tilde{V}$ to $c_{i-1}$:
15: \hspace{2em} for $v \in \text{neigh}(c_{i-1})$ do
16: \hspace{3em} $\hat{e} := \{c_{i-1}, v\}$
17: \hspace{3em} $H^{(J_e)} := (N_{c_{i-1}})_{\beta^e \rightarrow \beta^e} \boxtimes N_{c_{i-1}}$
18: \hspace{2em} end for
19: $\tilde{V} := \bigcup_{v \in \text{neigh}(c_{i-1})} \text{desc}_{c_{i-1}}(v)$
20: while $\tilde{V} \neq \text{leaves}(c_0)$ do
21: \hspace{2em} for $\tilde{v} \in \tilde{V} \setminus \text{leaves}(c_0)$ do
22: \hspace{3em} $p_{-1} := \text{pred}_{c_0}(\tilde{v})$, $\hat{e} := \{p_{-1}, \tilde{v}\}$
23: \hspace{3em} for $b \in \text{desc}_{c_{i-1}}(\tilde{v})$ do
24: \hspace{4em} $H^{(J_{\hat{e} \hat{e} \hat{e}})} := (N_b)_{\beta^e \rightarrow \beta^e \rightarrow \beta^e} \boxtimes H^{(J_e)} \boxtimes N_b$
25: \hspace{2em} end for
26: \hspace{2em} $\tilde{V} := \bigcup_{\tilde{v} \in \tilde{V}} \text{desc}_{c_{i-1}}(\tilde{v})$
27: \hspace{2em} end for
28: end while

// calculate $\{B^{(i-1)} \}_e \in K_{S_{i-1}}$ (cf. Proposition 6.6) or possibly limit set via maximal distance of $\tilde{v} \in \tilde{V}$ to $c_{i-1}$:
29: $\tilde{V} := \text{leaves}(c_0)$
30: while $\tilde{V} \neq \{c_0\}$ do
31: \hspace{2em} for $\tilde{v} \in \tilde{V} \setminus \{c_0\}$ do
32: \hspace{3em} $p_{-1} := \text{pred}_{c_0}(\tilde{v})$, $\hat{e} := \{p_{-1}, \tilde{v}\}$
33: \hspace{3em} for $b \in \text{desc}_{c_{i-1}}(\tilde{v})$ do
34: \hspace{4em} $B^{(i-1, J_{\hat{e} \hat{e} \hat{e}})} := (N_b)_{\beta^e \rightarrow \beta^e \rightarrow \beta^e \rightarrow \beta^e \rightarrow \beta^e} \boxtimes B^{(i-1, J_{\hat{e} \hat{e} \hat{e}})} \boxtimes \text{Netensor} \tilde{v}$
35: \hspace{3em} end for
36: \hspace{2em} end for
37: end while
\end{algorithm}
ASYMPTOTIC LOG-DET SUM-OF-RANKS MINIMIZATION VIA TENSOR (A)IRLS SM19

\[ B^{(i-1,J,\ell)} := (H^{(i-1,J,\ell)} + \gamma^{(i-1)} I)^{-1} + \sum_{b \in \text{desc}_{\ell-1}(\tilde{v})} \tilde{B}^{(J,\ell,b)} \]

\[ \tilde{V} := \bigcup_{\tilde{v} \in \tilde{V}} \text{pred}_{\ell}(\tilde{v}) \]

\[ \text{end while} \]

// solve and update representation (cf. (6.3) and Propositions 6.5 and 6.6) possibly using iterative solver (cf. Section SM6.5)

\[ A^{(i-1)} := \Delta_{\alpha_c, \alpha_e} \bigotimes_{v \in \text{neigh}(c_{i-1})} \left( \tilde{\bigotimes}_{c \in E_{c_{i-1}} \setminus \{c_{i-1}, v\}} \Delta_{\beta^e, \beta^c} \right) \bigotimes B^{(J_{c_{i-1}, v})} \]

\[ F^{(i-1)} := L_{c_{i-1}} \bigotimes_{v \in \text{neigh}(c_{i-1})} S^{(J_{c_{i-1}, v})} \]

\[ \text{solve } (F^{(i-1)}_{m_{c_{i-1}}} \rightarrow m'_{c_{i-1}} \bigotimes F^{(i-1)} + e^c \gamma^{(i-1)} A^{(i-1)} \bigotimes N_{c_{i-1}}^+ = F^{(i-1)}_{m_{c_{i-1}}} \rightarrow m'_{c_{i-1}} \bigotimes y} \]

\[ N_{c_{i-1}} := N_{c_{i-1}}^+ \]

// let \( J_\ell \in \mathcal{K}^{S_{\ell_i}}, e \in E \), and shift root:

\[ \text{set } c_i \in V \setminus \{c_{i-1}\} \quad (\text{cf. Section 6.1}) \]

// orthonormalize \( \{N_\ell\}_{\ell \in V} \) with respect to \( c_i \) (cf. Theorem 5.11) and supplement the missing branch evaluations of \( \{S^{(J_{\ell})}\}_{\ell \in E} \) (cf. Proposition 6.5)

\[ \text{set } p \text{ as path from including } c_{i-1} \text{ to including } c_i \quad (\text{cf. Section 5.2}) \]

\[ \text{for } j = 1, \ldots, |p| \text{ do } \]

\[ \tilde{c} := \{p_{j-1}, p_j\} \]

\[ U \text{ diag}(\sigma(J_\ell)) V^T := N^{[m_{p_{j-1}} \setminus \{\beta^\ell\}]}_{p_{j-1}} \]

\[ \text{possibly adapt rank } r \in \mathbb{N}^K \text{ via according modification of the SVD components } U, \text{ diag}(\sigma(J_\ell)) \text{ and } V^T \quad (\text{cf. Section SM6.3}) \]

\[ N^{[m_{p_{j-1}} \setminus \{\beta^\ell\}]}_{p_{j-1}} := U \text{ diag}(\sigma(J_\ell)), N^{[\beta^\ell]}_{p_{j-1}} := V^T N^{[\beta^\ell]}_{p_{j-1}} \]

\[ S^{(J_\ell)} := (L_{p_{j-1}} \bigotimes N^{[\beta^\ell]}_{p_{j-1}}) \bigotimes_{\tilde{h} \in \text{desc}_{\ell}(p_{j-1})} S^{(J_{p_{j-1}, \tilde{h}})}_{\ell} \]

\[ \text{end for} \]

// implicitly declare iterate \( X^{(i)} = \tau_r(\{N_\ell\}_{\ell \in V}) \)

\[ \text{set } \gamma^{(i)} \leq \gamma^{(i-1)} \quad (\text{possibly bound from above, cf. Section SM6.2}) \]

\[ \text{consider break possibly also based on validation residual} \]

\[ \text{end for} \]

\[ \text{possibly repeat as post-iteration with adapted parameters to appropriately truncate representation (cf. Section SM6.4)} \]
SM6. Practical and Heuristic Aspects

In Experiment 7.4, the AIRLS-0K algorithm is enhanced through the use of the following heuristics as embedded in Algorithm 3.

SM6.1. Validation set. A fraction of measurements is passively used to instead validate the progress allowing for more suitable breaking criteria and to adaptively control the parameter $\gamma$ (cf. Section SM6.2). This however assumes that the algorithm despite the decreased number of actively used measurements still converges to the essentially same solution.

SM6.2. Adaptive decay of regularization parameter $\gamma$. Practice shows that, additionally to the constant decline, carefully bounding $\gamma$ from above by a value proportional to the residual norm on the validation measurements (cf. Section SM6.1) can speed up convergence considerably without infringing upon the approximation.

SM6.3. Explicit rank adaption. The AIRLS-0K algorithm necessarily relies on the choice of some $\{r(J)\}_{J \in K}$ which bounds the ranks of the iterate. An adaptive determination can save a considerable amount of computational complexity. Introducing or removing a singular value $\sigma^{(J)}_{r(J)}(X)$ (thus changing the rank of the iterate), that is small compared to $\gamma$, only marginally influences the iteration. A method that has proven itself reliable in practice is to adapt each single rank $r(J)$, $J \in K$, such that always $\sigma^{(J)}_{r(J)-2}(X) > \frac{1}{2} \sqrt{\gamma}$, but $\sigma^{(J)}_{r(J)-1}(X) < \frac{1}{2} \sqrt{\gamma}$. Thereby, there are always exactly two comparatively low singular values with respect to each subset $J$.

SM6.4. AIRLS-0K internal post-iteration. In particular if the ranks are explicitly adapted, some singular values of the final iterate may be small enough such that a truncation of such seems more reasonable. Instead of a separate procedure that does not consider the original problem setting, a better approximation can be achieved by letting the algorithm proceed some additional iterations but with adapted meta parameters and for a specifically chosen value $\gamma$. Alternatively for small dimensions, the post-iteration scheme as discussed in Section 7.4 may be utilized.

SM6.5. Solving linear subsystems with iterative solvers. The linear subproblems that appear in each optimization step might become too large to solve explicitly using ordinary, full matrix vector calculus. Iterative solvers, such as preconditioned CG, can be applied to reduce the order of complexity significantly by exploiting the given low rank as well as additive structures. Whether this is truly beneficial naturally depends on the exact sizes that are involved, and not least the implementation.