Electronic Transactions on Numerical Analysis.
Volume 62, pp. 22–57, 2024.

ISSN 1068–9613.
DOI: 10.1553/etna_vol62s22

ETNA
Kent State University and
Johann Radon Institute (RICAM)

QUASI-ORTHOGONALIZATION FOR ALTERNATING NON-NEGATIVE TENSOR FACTORIZATION*

LARS GRASEDYCK†, MAREN KLEVER†, AND SEBASTIAN KRÄMER†

Abstract. Low-rank tensor formats allow for efficient handling of high-dimensional objects. In many applications, it is crucial to preserve the non-negativity in the approximation, for instance, by constraining all cores to be non-negative. Common alternating strategies reduce the high-dimensional problem to a sequence of low-dimensional subproblems but often suffer from slow convergence and persistence in local minima. In order to counteract this, we propose a new quasi-orthogonalization strategy as an intermediate step between the alternating minimization steps that preserves non-negativity. It allows one to improve the expressivity in each individual factor by modifying the current factorization within the equivalence class representing the same tensor.

Key words. non-negative factorization, orthogonalization, M-matrices, low-rank tensors, alternating least-squares, high-dimensional problems

AMS subject classifications. 15-06, 65F06, 65D40

1. Introduction.

Non-negative tensor factorization is a concept of decomposing or approximating a tensor with a set of smaller non-negative factors. Given a linear tensor operator $\mathcal{A} \in \mathbb{R}^{(n_1 \times \cdots \times n_d) \times (n_1 \times \cdots \times n_d)}$ and a right-hand side tensor $\mathbf{B} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, we address the problem to find $\mathbf{X}^\star = \tau_k((\mathbf{X}^\star_\mu)_\mu) \in \mathbb{R}^{n_1 \times \cdots \times n_d}_{\geq 0}$ with

$$(\mathbf{X}^\star_\mu)_\mu \in \operatorname*{argmin}_{(\mathbf{X}_\mu)_\mu} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 \quad \text{s.t. } \mathbf{X} = \tau_k((\mathbf{X}_\mu)_\mu),\ \mathbf{X}_\mu \geq 0 \ \forall\, \mu, \tag{1.1}$$

where $\|\cdot\|_F$ is the Frobenius norm, the inequalities are meant entry-wise, and $\tau_k(\cdot)$ is a multilinear factorization map which specifies a tree tensor format, such as the tensor-train [34] or hierarchical Tucker format [11, 13]. Here, k quantifies the sizes of the inputs of $\tau_k(\cdot)$ and $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_\mu) \geq 0$ is implied by $\mathbf{X}_\mu \geq 0$ for all μ.

The general problem (1.1) includes, in particular, the well-known non-negative matrix factorization problem: for a matrix $B \in \mathbb{R}^{n_1 \times n_2}$, that is, a two-dimensional ($d = 2$) tensor, find $X^\star = \tau_k(Y^\star, Z^\star) \in \mathbb{R}^{n_1 \times n_2}_{\geq 0}$ with

$$(Y^\star, Z^\star) \in \operatorname*{argmin}_{(Y,Z)} \|X - B\|_F^2 \quad \text{s.t. } X = \tau_k(Y, Z),\ Y, Z \geq 0, \tag{1.2}$$

where here $\tau_k(Y, Z) = YZ^T$ for $Y \in \mathbb{R}^{n_1 \times k}_{\geq 0}$ and $Z \in \mathbb{R}^{n_2 \times k}_{\geq 0}$ with Z^T denoting its transpose. Popular strategies to solve (1.1) and (1.2) are alternating optimization methods. In each micro-step for one ν, one fixes $\mathbf{X}_{\neq \nu} := (\mathbf{X}_\mu)_{\mu \neq \nu}$ and solves

$$\mathbf{X}^+_\nu \in \operatorname*{argmin}_{\mathbf{X}_\nu \geq 0} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 \quad \text{s.t. } \mathbf{X} = \tau_k(\mathbf{X}_{\neq \nu}, \mathbf{X}_\nu). \tag{1.3}$$

This allows one to reduce the large multilinear problem into a sequence of comparatively small, non-negative least-squares problems.

*Received November 29, 2023. Accepted June 16, 2024. Published online on July 3, 2024. Recommended by Peter Benner. This work was partially supported by the German Research Foundation (DFG) grant GR-3179/6-1 "Tensorapproximationsmethoden zur Modellierung von Tumorprogression" and SPP-1886 "Polymorphic uncertainty modelling for the numerical design of structures" under grant GR3179/5-2.

†Institute for Geometry and Applied Mathematics, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany ({lgr, klever, kraemer}@igpm.rwth-aachen.de), ORCID: 0000-0001-7572-2604, 0000-0003-2614-163X, and 0000-0001-9737-5906.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

1.1. Context and literature. High-dimensional problems with non-negativity constraints arise naturally in several areas, e.g., for probability distributions over combinatorial state spaces in stochastic automata networks [24, 35], queueing theory [4, 21], and chemical reaction networks [1, 27]. By exploiting certain structures of the problem, the curse of dimensionality can be broken using low-rank tensor formats such as the tensor-train (TT) or matrix product state (MPS) format [34, 44, 46, 53] or the hierarchical Tucker format [11, 13]. When utilizing such formats, one often does not obtain an exact representation, but a good approximation. In many applications, where the target tensor is non-negative, it is crucial to preserve this in the approximation. A particular class of such are probability distributions, since negative probabilities may not allow for an interpretation and should therefore be prohibited. Such a high-dimensional probability distribution, which occurs in tumor progression modeling [37], is considered in Section 4.3.

One way to deal with non-negativity conditions on a tensor $\mathbf{X}$ are so-called *non-negative low-rank approximations* $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_\mu)$, where the individual factors $\mathbf{X}_\mu$ may have negative entries and only the represented tensor satisfies $\mathbf{X} \geq 0$. This approach is less restrictive with respect to the choice of $(\mathbf{X}_\mu)_\mu$, but fulfilling $\mathbf{X} \geq 0$ can be restrictively hard. To deal with this problem, alternating projections between a low-rank manifold and the non-negative orthant $\mathbb{R}_{\geq 0}^{n_1 \times \cdots \times n_d}$ may be used. These have recently been extended from matrices [39] to higher-dimensional tensors [38, 40]. However, even determining if a high-dimensional tensor $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_\mu) \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ in a low-rank format satisfies $\mathbf{X} \geq 0$ without computing all n^d entries is a non-trivial task. In practice, these methods can oftentimes reduce the number and absolute value of negative entries. However, there is no guarantee that the resulting low-rank tensor is completely non-negative. As an alternative to alternating projections, there are also recent works on approximating non-negative objects with low rank based on squaring, i.e., using $\mathbf{X} = (\tau_k((\mathbf{X}_\mu)_\mu))^2$ (with component-wise squaring); see [28]. This naturally guarantees the non-negativity of $\mathbf{X}$. However, minimizing $\|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2$ over $\mathbf{X} = (\tau_k((\mathbf{X}_\mu)_\mu))^2$ is not multilinear any more, which may lead to other issues.

A different way to combine both low-rank tensor formats and non-negativity is *non-negative tensor factorization*, where all factors are themselves non-negative, i.e., $\mathbf{X}_\mu \geq 0$ for all μ. For matrices, the problem is known as *non-negative matrix factorization*. This approach is more restrictive, but directly guarantees that $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_\mu)$ is non-negative. In addition, the individual non-negative factors $\mathbf{X}_\mu$ themselves can be interpreted in some applications; for instance, in the context of graphical probability models [36].

For ordinary low-rank tensor formats, there exist many techniques to solve a linear system $\mathcal{A}(\mathbf{X}) = \mathbf{B}$, including iterative and optimization methods; see, e.g., [12] for an overview. As arithmetic operations typically lead to an increase in the representation ranks of tensors, applying iterative solvers within low-rank formats essentially relies on a quasi-optimal truncation, which allows one to reduce the ranks in an error-controlled way [11, 13, 34]. For non-negative tensor factorization, this boils down to solving (1.1) when $\mathcal{A}$ is the identity, which is non-trivial even beyond quasi-optimality. There are also strategies that rely on constructions of suitable manifolds, requiring the identification of equivalence classes of factorizations that describe the same tensor; see, e.g., [7, 16, 43] for classical low-rank formats. Without non-negativity constraints on the factors, the degrees of freedom are given by regular transfer matrices between two neighboring factors each. However, for non-negative factorizations, this is not straightforward.

In the context of non-negative matrix factorization (1.2), there are also methods that use certain non-negative functions to describe the entries of the factors. When the target matrix B contains samples of non-negative smooth (or mostly smooth) functions, formulating the coefficients of Y and/or Z as certain discretizations of continuous non-negative functions such

ETNA
Kent State University and
Johann Radon Institute (RICAM)

as polynomials [6], splines [50], or rational functions [14] can be beneficial, as they allow for taking prior, problem-specific information into account.

In the context of non-negative tensor factorizations, there are also optimization approaches that operate on the entire parameter space; e.g., [45]. This allows one to use more general methods, for instance, Gauss–Newton-like methods at moderate dimensions [45]. For higher-dimensional problems, oftentimes alternating optimization strategies are used, splitting (1.1) into a sequence of micro-steps (1.3). The question of how to solve (1.3) has been studied extensively in different communities. These include, for instance, multiplicative updates [26], hierarchical alternating least-squares [51], alternating direction method of multipliers [18] related to non-negative matrix factorization, or more general interior-point methods [48] for quadratic programs; see, e.g., [5], [22, Section 5.6], and [10, Chapter 4] for overviews. Extending vanilla alternating non-negative strategies, further acceleration and extrapolation methods are developed in order to improve (empirical) convergence speed for alternating non-negative matrix and tensor factorization; see, e.g., [47, Section 3.4] as well as [29, 31] for some recent works.

1.2. Motivation. In this work, we address the issue that alternating non-negative tensor factorization often experiences slow convergence and persistence in local minima [10, Chapter 3 and p. 169 item 1], especially for higher dimensions. One property related to alternating procedures is the expressivity we define as follows.

DEFINITION 1.1 (Expressivity). *Let $\nu \in [d]$, $\mathbf{X}_{\neq\nu} := (\mathbf{X}_\mu)_{\mu\neq\nu}$, and sizes k be given. Then the expressivity in the ν-th factor given $\mathbf{X}_{\neq\nu}$* (without *non-negativity constraints) is defined as the range of* $\mathbb{R}^{\operatorname{sizes}(\mathbf{X}_\nu)} \to \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$, $\mathbf{X}_\nu \mapsto \tau_k(\mathbf{X}_{\neq\nu}, \mathbf{X}_\nu)$.

Given $\mathbf{X}_{\neq\nu} = (\mathbf{X}_\mu)_{\mu\neq\nu} \geq 0$, the expressivity in the ν-th factor with *non-negativity constraints is the range of* $\mathbb{R}_{\geq 0}^{\operatorname{sizes}(\mathbf{X}_\nu)} \to \mathbb{R}_{\geq 0}^{n_1\times n_2\times\cdots\times n_d}$, $\mathbf{X}_\nu \mapsto \tau_k(\mathbf{X}_{\neq\nu}, \mathbf{X}_\nu)$.

Please note that this expressivity is independent of $\mathcal{A}$ and $\mathbf{B}$ in (1.3) and

$$\mathbf{X}_\nu^+ \in \operatorname*{argmin}_{\mathbf{X}_\nu} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 \quad \text{s.t. } \mathbf{X} = \tau_k(\mathbf{X}_{\neq\nu}, \mathbf{X}_\nu),$$

respectively. For this reason, we assign the expressivity to a fixed factor (given all other factors) instead of the micro-step, although they are also related.

In contrast to the classical setting without non-negativity constraints, the expressivity in each factor ν given $\mathbf{X}_{\neq\nu} = (\mathbf{X}_\mu)_{\mu\neq\nu} \geq 0$ strongly depends on the choice of the fixed factors $\mathbf{X}_{\neq\nu}$ and also on the particular representative of the equivalence class describing the same tensor $\mathbf{X}$ as we will discuss in the following. For this reason, we want to maximize this expressivity in each intermediate step between consecutive micro-steps (1.3).

By reshaping the factor $\mathbf{X}_\nu$ and the difference $\mathcal{A}(\mathbf{X}) - \mathbf{B}$ into appropriate matrices, the micro-step (1.3) behaves similarly to those used in non-negative matrix factorization,

$$Z^+ \in \operatorname*{argmin}_{Z\in\mathbb{R}^{n_2\times k}} \|X - B\|_F^2 \quad \text{s.t. } X = \tau_k(Y, Z),\ Z \geq 0,$$

for fixed $Y \in \mathbb{R}_{\geq 0}^{n_1\times k}$ (and analogously for Y^+ with fixed $Z \in \mathbb{R}_{\geq 0}^{n_2\times k}$).

For the classical low-rank matrix approximation, i.e., (1.2), without non-negativity constraints on $Y \in \mathbb{R}^{n_1\times k}$ and $Z \in \mathbb{R}^{n_2\times k}$, orthogonalization via QR decomposition is used as an intermediate step. By replacing Y with Q from its reduced QR decomposition $Y = QR$, the set of matrices that can be obtained for fixed Y remains unchanged:

$$\begin{aligned}\{X \in \mathbb{R}^{n_1\times n_2} \colon \operatorname{range}(X) \subseteq \operatorname{range}(Y)\} &= \{X = \tau_k(Y, Z) \colon Z \in \mathbb{R}^{n_2\times k}\} \\ &= \{X = \tau_k(Q, W) \colon W \in \mathbb{R}^{n_2\times \operatorname{rank}(Y)}\} = \{X \in \mathbb{R}^{n_1\times n_2} \colon \operatorname{range}(X) \subseteq \operatorname{range}(Q)\}.\end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

This means that the expressivities in the factor Z given $Y \in \mathbb{R}^{n_1 \times k}$ or $Q \in \mathbb{R}^{n_1 \times \operatorname{rank}(Y)}$, respectively, are equal, with $\operatorname{range}(Y) = \operatorname{range}(Q)$. When restricting $Y, Z \geq 0$, the expressivity in the factor Z given $Y \geq 0$ depends on the non-negative range, $\operatorname{range}_{\geq 0}(Y) := \{Yz \colon z \in \mathbb{R}^k_{\geq 0}\}$, spanned by the columns in Y; see Definition 3.1. For Y without redundant columns, any non-negative factorization $Y = VN$ with $V, N \geq 0$ and a square N that is not a permuted diagonal matrix leads to $\operatorname{range}_{\geq 0}(Y) \subsetneq \operatorname{range}_{\geq 0}(V)$; see Theorem 3.10. Thus, the expressivity in the factor Z given $Y \geq 0$ relies not only on the choice for Y, but also on the representative of the equivalence class representing $\mathbf{X} = \tau_k(Y, Z)$.

EXAMPLE 1.2. Let $n_1 = 3$, $k = 2$, and

$$Y := \begin{bmatrix} 1 & 1/2 \\ 2 & 1 \\ 3/2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} =: VN.$$

Then $V, N \geq 0$, $\operatorname{rank}(N) = 2$, and $\{\tau_k(Y, Z) \colon Z \geq 0\} \subsetneq \{\tau_k(V, W) \colon W \geq 0\}$ since

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \in \operatorname{range}_{\geq 0}(V) \setminus \operatorname{range}_{\geq 0}(Y).$$

Thus, switching to the representation $\mathbf{X} = \tau_k(V, ZN^T)$ instead of $\mathbf{X} = \tau_k(Y, Z)$ would be preferable for an update of the second factor.

Similar holds true for higher-dimensional tensors. For instance, let $\mathbf{X} = \tau_k(\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3) \in \mathbb{R}^{n \times n \times n}_{\geq 0}$ with $\mathbf{Y}_1, \mathbf{Y}_3^T \in \mathbb{R}^{n \times k}_{\geq 0}$ and $\mathbf{Y}_2 \in \mathbb{R}^{k \times n \times k}_{\geq 0}$. Then one can modify the factors $(\mathbf{Y}_\mu)_{\mu \in [3]}$ by regular matrices $N_1, N_2 \in \mathbb{R}^{k \times k}_{\geq 0}$ such that for all $i \in \{1, \dots n\}$ it holds

$$\mathbf{V}_1 := \mathbf{Y}_1 N_1^{-1} \geq 0, \quad \mathbf{V}_3 := N_2 \mathbf{Y}_3 \geq 0 \quad \text{and} \quad (\mathbf{V}_2)_{:,i,:} := N_1 (\mathbf{Y}_2)_{:,i,:} N_2^{-1} \geq 0 \; \forall \, i \in [n]$$

without changing the represented tensor, that is, $\mathbf{X} = \tau_k(\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3)$. Again, in contrast to classical low-rank approximation without non-negativity constraints, the expressivity in the ν-th factor given $\mathbf{Y}_{\neq \nu} \geq 0$ and $\mathbf{V}_{\neq \nu} \geq 0$, respectively, that is, the ranges of $\mathbf{X}_\nu \mapsto \tau_k(\mathbf{Y}_{\neq \nu}, \mathbf{X}_\nu)$ and $\mathbf{X}_\nu \mapsto \tau_k(\mathbf{V}_{\neq \nu}, \mathbf{X}_\nu)$ for $\mathbf{X}_\nu \geq 0$ (see Definition 1.1), can be different.

Ordinary orthogonalization in general violates the non-negativity conditions for the individual factors and thus cannot directly be used here. One common alternative is to simply rescale the factors by using diagonal transfer matrices whose diagonal entries are all positive; see Algorithm 2.1. This may help to improve stability, but does not allow for further improvements with respect to the expressivity as in Example 1.2.

1.3. Main contribution. Based on the fact that the expressivity in each factor for non-negative factorization relies on the degrees of freedom between two factors (see Example 1.2), our main contribution is the development of a *quasi-orthogonalization*. Its aim is to maximize the non-negative range of each fixed factor while preserving the non-negativity of all factors. We illustrate this idea in Examples 3.14 and 3.15. In Theorem 3.5, we prove that there exists a minimal subset of columns that generates the non-negative range. For this reason, all other columns can either be removed or replaced without decreasing its non-negative range. Theorem 3.10 shows that, for Y consisting only of such generators, strictly increasing its non-negative range is equivalent to finding a non-negative factorization $Y = VN$ with square N not being a permuted diagonal matrix. One way to obtain this factorization is to find an inverse non-negative transfer matrix such that its application to the current factor retains non-negativity. Here, we focus on a particular subset of M-matrices; see Notation 3.18. We

ETNA
Kent State University and
Johann Radon Institute (RICAM)

prove that these matrices are inverse non-negative under mild conditions; see Theorems 3.20 and 3.29. In Section 3.4 we introduce a method for finding an appropriate matrix through modifications of Y in a column-wise manner. Theorem 3.32 proves the equivalence of an increase in the non-negative range by such column-wise modification and the existence of such an M-matrix. This allows us to formulate the quasi-orthogonalization in Algorithm 3.1, which has negligible effort compared to solving a micro-step; see Lemma 3.39 and Remark 3.41.

1.4. Related methods and motivation thereof. One work we would like to highlight here is [9]: there, a preprocessing strategy for non-negative matrix factorization is proposed which is based on similar observations. Given a target matrix $B \in \mathbb{R}_{\geq 0}^{n_1 \times n_2}$, their goal is to obtain a more well-posed non-negative matrix factorization problem for BM compared to B via an M-matrix $M \in \mathbb{R}^{n_2 \times n_2}$. The transfer matrix M is chosen such that M is inverse non-negative and $BM \geq 0$ is sparse. A set of specific M-matrices, which is equivalent to $\mathcal{M}_B$ in Notation 3.18, is selected, and conditions for the non-negativity of the inverses are derived. To ensure sparsity, the Frobenius norm of BM as an approximation of its ℓ_0-norm is minimized. Reference [9] also provides a geometric interpretation of this preprocessing using some examples, which is consistent with our idea of maximizing the non-negative range. In contrast to [9], we are interested, in particular, in the behavior of the non-negative range by changing to representative of the equivalence class of a non-negative factorization.

1.5. Organization of the remainder of the paper. An overview of the notation and preliminary concepts is given in Section 2, including a brief introduction to the tensor-train format in Section 2.1 and the alternating non-negative tensor factorization in Section 2.2. In Section 3 we derive the quasi-orthogonalization strategy and provide an analysis of it. Section 4 contains detailed numerical experiments on the quasi-orthogonalization for the non-negative factorization of symmetric polynomials in Section 4.2 and certain probability distributions for high-dimensional Markov chains in Section 4.3. We conclude in Section 5.

2. Notation and preliminary concepts. For the remainder of this paper, we use the following notation and concepts. We write $[n] := \{1, \dots, n\}$ for any $n \in \mathbb{N}$ and $\mathbb{R}_{\geq 0} := [0, \infty)$. For a matrix $Y \in \mathbb{R}^{n_1 \times k}$, we denote its entry at (i, j) by $Y_{i,j}$, its i-th row by $Y_{i,:}$, and its j-th column by $Y_{:,j}$. The matrix without the i-th column (or row) is denoted by $Y_{:,\neq i}$ (or $Y_{\neq i,:}$), the set of its columns by $\operatorname{col}(Y) := \{Y_{:,\ell} \colon \ell \in [k]\}$, and its range by $\operatorname{range}(Y) := \{Y\lambda \colon \lambda \in \mathbb{R}^k\}$. A d-dimensional tensor with mode sizes $n_1, \dots, n_d \in \mathbb{N}$ is an object $\mathbf{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. We write $\mathbf{X}_i := \mathbf{X}_{i_1,\dots,i_d}$ for the evaluation of $\mathbf{X}$ at $i := (i_1, \dots, i_d)$. Inequalities are always meant entry-wise, e.g., $\mathbf{X} \geq 0$ if and only if $\mathbf{X}_i \geq 0$ for all i. We denote the Frobenius norm of $\mathbf{X}$ by $\|\mathbf{X}\|_F := \sqrt{\sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \mathbf{X}_{i_1,\dots,i_d}^2}$. We write $e_i \in \{0,1\}^n$ for the i-th canonical unit tensor, that is, $(e_i)_j = 1$ if $j = i$ and $(e_i)_j = 0$ otherwise, and $\mathrm{Id}_n \in \mathbb{R}^{n \times n}$ for the identity operator. Further, $\mathbf{0}_n$ and $\mathbf{1}_n \in \mathbb{R}^n$ denote the tensors of all zeros and ones, respectively. We write $\operatorname{diag}(x_1, \dots, x_n) \in \mathbb{R}^{n \times n}$ for the diagonal matrix with diagonal entries $x_1, \dots, x_n$.

2.1. Low-rank tensor formats: tensor-train format. We briefly introduce the concept and notation of low-rank tensor formats. A basic idea, on which most tensor formats are based, is to unfold a high-dimensional tensor into matrices by partitioning its modes, and reshaping the tensor accordingly; see, e.g., [22, Section 2.4].

DEFINITION 2.1 (Unfolding). *The unfolding of a tensor $\mathbf{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ with respect to the row modes $\gamma \subseteq [d]$ and column modes $\gamma^c := [d] \setminus \gamma$ is the matrix $\mathbf{X}^{(\gamma)} \in \mathbb{R}^{\prod_{\mu \in \gamma} n_\mu \times \prod_{\nu \in \gamma^c} n_\nu}$ with*

$$(\mathbf{X}^{(\gamma)})_{(i_\mu)_{\mu \in \gamma}, (i_\nu)_{\nu \in \gamma^c}} := \mathbf{X}_i$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

for all indices $i = (i_\mu)_{\mu\in[d]}$. *In particular,* $\mathbf{X}^{([d])}$ *is the vectorization of* $\mathbf{X}$ *as a column and* $\mathbf{X}^{(\emptyset)}$ *as a row.*

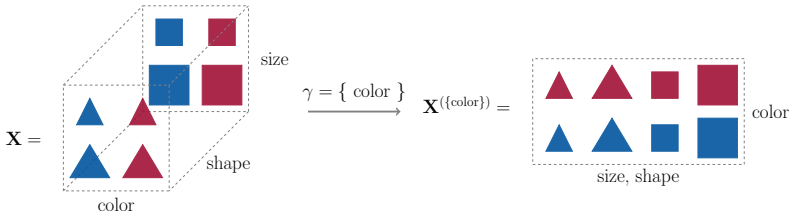


FIG. 2.1. *Unfolding of a three-dimensional tensor* $\mathbf{X}$ *in size* $\times$ *color* $\times$ *shape into a matrix* $\mathbf{X}^{(\gamma)}$ *with* $\gamma = \{color\}$.

Figure 2.1 shows a schematic example for an unfolding of a three-dimensional tensor into a matrix. This approach allows one to transfer several concepts and strategies from matrices back to tensors. For simplicity, we focus on the tensor-train format [34], also known as MPS format [44, 46]. However, all concepts presented can also be transferred to more general hierarchical Tucker [11, 13] or tree tensor formats.

DEFINITION 2.2 (Tensor-train representation). *A tensor* $\mathbf{X} \in \mathbb{R}^{n_1\times\cdots\times n_d}$ *has a TT representation of size* $k := (k_0, k_1, \dots, k_d) \in \mathbb{N}^{d+1}$ *with* $k_0 = k_d = 1$ *if and only if there exist* $(\mathbf{X}_\mu)_{\mu\in[d]}$, *called TT cores, with* $\mathbf{X}_\mu \in \mathbb{R}^{k_{\mu-1}\times n_\mu\times k_\mu}$, *such that*

$$\mathbf{X}_i = \sum_{\ell_1=1}^{k_1} \cdots \sum_{\ell_{d-1}=1}^{k_{d-1}} (\mathbf{X}_1)_{1,i_1,\ell_1} (\mathbf{X}_2)_{\ell_1,i_2,\ell_2} \cdots (\mathbf{X}_d)_{\ell_{d-1},i_d,1}$$

for all $i = (i_\mu)_{\mu\in[d]}$. *In this case, we call* $\mathbf{X}$ *a TT tensor and identify* $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]})$.

Similarly, we define a tensor-train operator as follows.

DEFINITION 2.3 (Tensor-train operator). *A linear operator* $\mathcal{A}$ *from* $\mathbb{R}^{n_1\times\cdots\times n_d}$ *to* $\mathbb{R}^{m_1\times\cdots\times m_d}$ *has a TT representation of size* $K := (K_0, K_1, \dots, K_d) \in \mathbb{N}^{d+1}$ *with* $K_0 = K_d = 1$ *if and only if there exist TT cores* $(\mathbf{A}_\mu)_{\mu\in[d]}$ *with* $\mathbf{A}_\mu \in \mathbb{R}^{K_{\mu-1}\times m_\mu\times n_\mu\times K_\mu}$ *such that*

$$\mathcal{A}_{i,j} = \sum_{\ell_1=1}^{K_1} \cdots \sum_{\ell_{d-1}=1}^{K_{d-1}} (\mathbf{A}_1)_{1,i_1,j_1,\ell_1} (\mathbf{A}_2)_{\ell_1,i_2,j_2,\ell_2} \cdots (\mathbf{A}_d)_{\ell_{d-1},i_d,j_d,1}$$

for all $i = (i_\mu)_{\mu\in[d]}$ *and* $j = (j_\mu)_{\mu\in[d]}$. *In this case, we call* $\mathcal{A}$ *a TT operator and identify* $\mathcal{A} = \tau_K((\mathbf{A}_\mu)_{\mu\in[d]})$.

REMARK 2.4. The TT tensors of size $k = (k_0, k_1, \dots, k_d)$ in Definition 2.2 are exactly the tensors $\mathbf{X}$ whose unfoldings $\mathbf{X}^{(\{1,\dots,\mu\})}$ have ranks of at most k_μ [34], i.e., $\operatorname{rank}(\mathbf{X}^{(\{1,\dots,\mu\})}) \le k_\mu$ for all $\mu \in [d]$. However, as we are interested in non-negative factorizations, these rank conditions are less important here.

Analogously to Definition 2.2, we define a non-negative tensor-train factorization as follows.

DEFINITION 2.5 (Non-negative tensor-train factorization). *A TT representation* $(\mathbf{X}_\mu)_{\mu\in[d]}$ *is called a non-negative factorization if and only if all TT cores are non-negative, i.e.,* $\mathbf{X}_\mu \ge 0$ *for all* $\mu \in [d]$.

The main benefits of the tensor-train format are its low storage and computational complexity for performing, for instance, the application of TT operators to TT tensors, inner

ETNA
Kent State University and
Johann Radon Institute (RICAM)

TABLE 2.1
Operations and their costs for TT tensors $\mathbf{X}$ and $\mathbf{C}$ and a TT operator $\mathcal{A}$ each of dimension d with constant mode sizes n and ranks component-wise bounded by k and K for $\mathcal{A}$, respectively, [32].

Operation	Formula	Cost
Storage		$\mathcal{O}(dnk^2)$
Evaluation	$\mathbf{X}_i$	$\mathcal{O}(dnk^2)$
Inner product	$\langle \mathbf{X}, \mathbf{C} \rangle$	$\mathcal{O}(dnk^3)$
Operator application	$\mathcal{A}(\mathbf{X})$	$\mathcal{O}(dn^2K^2k^2)$

products between, and evaluations of such [34]. We summarize some of these operations and their respective cost in Table 2.1. In addition to this formal definition of TT tensors, we also use a graphical representation of such, so-called tensor networks; similar to, for instance, [17]. Here, individual d-dimensional tensors are represented as nodes with d legs, i.e., half-edges connected to only one node. Figure 2.2(a) shows some examples of d-dimensional tensors. Contractions of two tensors over a certain mode are represented as a common edge between their respective tensor nodes.

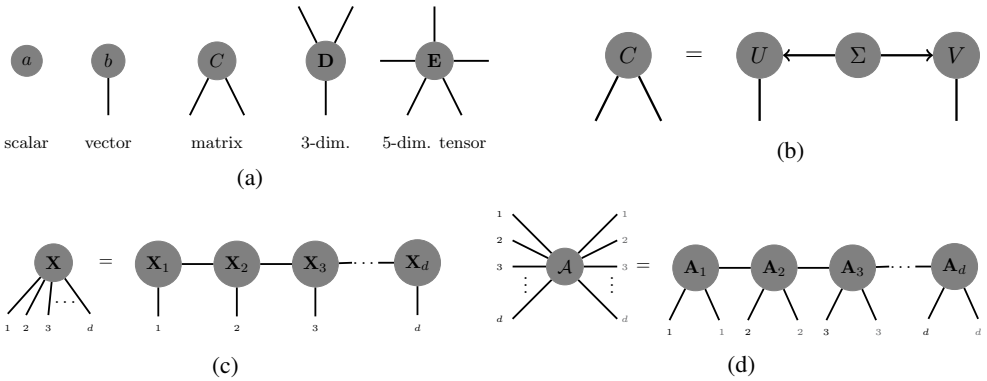


FIG. 2.2. *Tensor node networks of (a) d-dimensional tensors with $d \in \{0, \dots, 5\}$, (b) the singular value decomposition $C = U\Sigma V^T = \tau_{\mathrm{rank}(C)}(U, \Sigma, V)$ with orthogonality represented by arrows, (c) a TT tensor $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu \in [d]})$, and (d) a TT operator $\mathcal{A} = \tau_K((\mathbf{A}_\mu)_{\mu \in [d]})$.*

2.2. Alternating non-negative tensor factorization. Motivated by the alternating least-squares strategies for the tensor-train approximation, we introduce alternating non-negative tensor factorization in the tensor-train format. If $\mathcal{A}$ and $\mathbf{B}$ are given in the TT format, the classical minimization problem

$$(\mathbf{X}^\star_\mu)_{\mu \in [d]} \in \operatorname*{argmin}_{(\mathbf{X}_\mu)_{\mu \in [d]}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 \quad \text{s.t. } \mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu \in [d]}) \tag{2.1}$$

can be solved using alternating least-squares; e.g., [17]. One solves the micro-steps

$$\mathbf{X}^+_\nu \in \operatorname*{argmin}_{\mathbf{X}_\nu \in \mathbb{R}^{k_{\nu-1} \times n_\nu \times k_\nu}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 \quad \text{s.t. } \mathbf{X} = \tau_k(\mathbf{X}_{\neq \nu}, \mathbf{X}_\nu) \tag{2.2}$$

for each $\nu \in [d]$ while fixing all others $\mathbf{X}_{\neq \nu} = (\mathbf{X}_\mu)_{\mu \neq \nu}$ alternately. Before solving (2.2), the current tensor is typically orthogonalized with respect to the core $\mathbf{X}_\nu$ using QR decompositions of the unfoldings $\mathbf{X}_\mu{}^{(\{1,2\})} = Q_\mu R_\mu$ for $\mu < \nu$ and $\mathbf{X}_\mu{}^{(\{1\})} = R_\mu Q_\mu$ for $\mu > \nu$. Starting at

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$\mu = 1$, one decomposes $\mathbf{X}_\mu^{(\{1,2\})} = Q_\mu R_\mu$ and moves R_μ to the next core by $\mathbf{X}_\mu^{(\{1,2\})} \leftarrow Q_\mu$ and $\mathbf{X}_{\mu+1}^{(\{1\})} \leftarrow R_\mu \mathbf{X}_{\mu+1}^{(\{1\})}$. Next, one moves on to the next core $\mu + 1$ until one reaches $\nu - 1$ (and analogously for $\eta = d$ to $\nu + 1$ in the opposite direction). Owing to the specific structure of the TT format, running the cores from $\nu = 1$ to d and back allows for a more efficient computation of the orthogonalization steps, as the QR decomposition only needs to be computed for the new updated core. Running one way, i.e., from $\nu = 1$ to d *or* back, solving the corresponding micro-steps (2.2) is commonly referred to as a *half-sweep*, whereas running it both ways, i.e., from $\nu = 1$ to d *and* back, is called a *sweep*.

The alternating procedure for solving (2.1) can easily be adapted to the non-negative case:

$$(\mathbf{X}_\mu^\star)_{\mu\in[d]} \in \operatorname*{argmin}_{(\mathbf{X}_\mu)_{\mu\in[d]}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 \quad \text{s.t. } \mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]}),\ \mathbf{X}_\mu \geq 0\ \forall\, \mu \in [d]. \tag{2.3}$$

To do so, instead of (2.2), one alternately solves the non-negative micro-step

$$\mathbf{X}_\nu^+ \in \operatorname*{argmin}_{\mathbf{X}_\nu \in \mathbb{R}^{k_{\nu-1}\times n_\nu \times k_\nu}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 \quad \text{s.t. } \mathbf{X} = \tau_k(\mathbf{X}_{\neq\nu}, \mathbf{X}_\nu),\ \mathbf{X}_\nu \geq 0 \tag{2.4}$$

for $\nu \in [d]$ and each fixed $\mathbf{X}_{\neq\nu}$. Each non-negative micro-step (2.4) is a non-negative least-squares problem, which can be solved using one of several available methods; see, e.g., [5] for an overview.

As mentioned before, the orthogonalization steps via QR decompositions cannot be used any more, as their results typically have negative entries. One common ansatz to improve the stability of the non-negative micro-step (2.4) and still keep all cores non-negative is to rescale all cores $(\mathbf{X}_\mu)_{\mu\neq\nu}$ such that

$$\|(\mathbf{X}_\mu^{(\{1,2\})})_{:,j}\| = \|(\mathbf{X}_\eta^{(\{1\})})_{\ell,:}\| = 1 \quad \forall\, j, \ell \text{ and } \forall\, \mu < \nu < \eta \tag{2.5}$$

for a given vector norm $\|\cdot\|$. This column-wise normalization can be performed using diagonal transfer matrices with positive diagonal elements, as formulated in Algorithm 2.1 and illustrated in Figure 2.3.

Algorithm 2.1: `diag` of $\mathbf{X}$ w.r.t. $\mathbf{X}_\nu$

Input: $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]})$ TT tensor with $\mathbf{X}_\mu \geq 0$, $\nu \in [d]$, $\|\cdot\|$ vector norm
Output: $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]})$ fulfilling (2.5)

for $\mu = 1, \dots, \nu - 1$ **do**
  $D_\mu := \operatorname{diag}(\|(\mathbf{X}_\mu^{(\{1,2\})})_{:,\ell}\| : \ell \in [k_\mu])$
  $\mathbf{X}_\mu^{(\{1,2\})} \leftarrow \mathbf{X}_\mu^{(\{1,2\})} D_\mu^{-1}$
  $\mathbf{X}_{\mu+1}^{(\{1\})} \leftarrow D_\mu \mathbf{X}_{\mu+1}^{(\{1\})}$
end
for $\eta = d, \dots, \nu + 1$ **do**
  $D_\eta := \operatorname{diag}(\|(\mathbf{X}_\eta^{(\{1\})})_{\ell,:}\| : \ell \in [k_{\eta-1}])$
  $\mathbf{X}_\eta^{(\{1\})} \leftarrow D_\eta^{-1} \mathbf{X}_\eta^{(\{1\})}$
  $\mathbf{X}_{\eta-1}^{(\{1,2\})} \leftarrow \mathbf{X}_{\eta-1}^{(\{1,2\})} D_\eta$
end
return $\mathbf{X}$

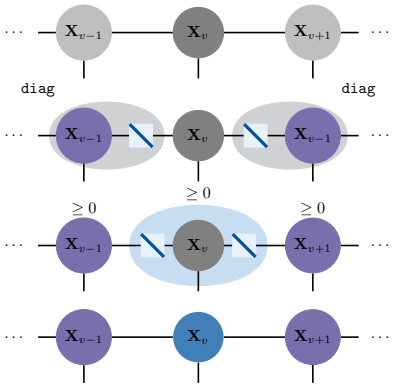


FIG. 2.3. *Illustration of* `diag` *for a non-negative TT tensor* $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]})$ *w.r.t.* $\mathbf{X}_\nu$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Besides stability issues, the expressivity in each non-negative factor ν given $\mathbf{X}_{\neq\nu} = (\mathbf{X}_\mu)_{\mu\neq\nu} \geq 0$, i.e., the range of $\mathbf{X}_\nu \mapsto \tau_k(\mathbf{X}_{\neq\nu}, \mathbf{X}_\nu)$, strongly depends on the representation chosen for the fixed cores $\mathbf{X}_{\neq\nu}$ (cf. Example 1.2) resulting in slow convergence and stagnation, in particular for higher dimensions. Therefore, our idea is to derive a quasi-orthogonalization strategy that maximizes the expressivity in each factor while preserving the non-negativity of all cores and not changing the represented tensor.

Given such a strategy, we summarize the resulting alternating non-negative least-squares method for non-negative tensor trains in Algorithm 2.2 and illustrate the quasi-orthogonalization step in Figure 2.4. To distinguish quasi-orthogonality from classical orthogonality, the arrow heads correspond to small vertical lines.

Algorithm 2.2: `TT-ANLS` for (2.3)

Input: $\mathcal{A}$ TT operator, $\mathbf{B}$ TT tensor, $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]})$ initial TT tensor with $\mathbf{X}_\mu \geq 0$
Output: $\mathbf{X}$ approximate solution of (2.3)

while *stop criteria are not fulfilled* **do**
  /* Half sweep $\nu = 1$ to d */
  for $\nu = 1, \ldots, d-1$ **do**
    Quasi-orthogonalize or normalize $\mathbf{X}$ w.r.t. $\mathbf{X}_\nu$
    Solve (2.4) and update $\mathbf{X}_\nu$
  end
  /* Half sweep $\nu = d$ to 1 */
  for $\nu = d, \ldots, 2$ **do**
    Quasi-orthogonalize or normalize $\mathbf{X}$ w.r.t. $\mathbf{X}_\nu$
    Solve (2.4) and update $\mathbf{X}_\nu$
  end
end
return $\mathbf{X}$

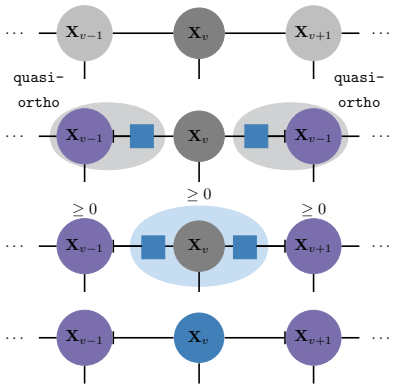


FIG. 2.4. *Illustration of* `quasi-ortho` *for* $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]})$ *w.r.t.* $\mathbf{X}_\nu$.

Besides the vanilla version in Algorithm 2.2, also other alternating strategies for non-negative tensor factorization can benefit from quasi-orthogonalization. As such an example, we extended the so-called heuristic extrapolation with restarts (HER) strategy from [29] to our problem (2.3) using tensor trains and an additional quasi-orthogonalization step. The resulting method is summarized in Algorithm A.1 and applied in Section 4 for comparison. For further details on HER, we refer the reader to [29].

3. Quasi-orthogonalization strategy.

For simplicity, we focus on the classical non-negative matrix factorization problem in (1.2) to derive our quasi-orthogonalization strategy. Similar to the classical orthogonalization, also the quasi-orthogonalization can easily be adapted to higher-dimensional tensors as explained in Remark 3.38 for the TT format.

3.1. Non-negative range of a non-negative matrix.

We start with some elementary definitions and observations motivating the quasi-orthogonalization. We recall the definition of conical combinations, the non-negative range of a non-negative matrix according to [15, pp. 101–102] and define an extreme column (analog to an extreme ray).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

DEFINITION 3.1 (Conical combination and non-negative range). *A conical combination* $c \in \mathbb{R}^{n_1}_{\geq 0}$ *of non-negative vectors* $y_1, \dots y_k \in \mathbb{R}^{n_1}_{\geq 0}$ *is a linear combination with non-negative coefficients, i.e.,*

$$c = \sum_{\ell=1}^{k} \lambda_\ell y_\ell \quad \text{with } \lambda_\ell \geq 0 \; \forall \, \ell \in [k].$$

The non-negative range of a non-negative matrix $Y \in \mathbb{R}^{n_1 \times k}_{\geq 0}$ *is defined as the set of all conical combinations of its columns:*

$$\operatorname{range}_{\geq 0}(Y) := \left\{ \sum_{\ell=1}^{k} \lambda_\ell Y_{:,\ell} \colon \lambda_\ell \geq 0 \; \forall \, \ell \in [k] \right\} = \{ Y\lambda \colon \lambda \in \mathbb{R}^k_{\geq 0} \}.$$

A column $Y_{:,\ell}$ *is called extreme if and only if it is not a conical combination of all others, i.e.,* $Y_{:,\ell} \notin \operatorname{range}_{\geq 0}(Y_{:,\neq \ell})$.

Note that the non-negative range of Y is a cone; in particular, it is the conical hull of $\operatorname{col}(Y)$. For fixed $Y \geq 0$, all elements in the range of $Z \mapsto \tau_k(Y, Z)$ for $Z \geq 0$ are column-wise contained in $\operatorname{range}_{\geq 0}(Y)$, i.e., $\operatorname{col}(\tau_k(Y, Z)) \subseteq \operatorname{range}_{\geq 0}(Y)$ for all $Z \geq 0$. In contrast to the classical low-rank approximation problem, the range of $Z \mapsto \tau_k(Y, Z)$ for $Z \geq 0$ strongly depends on the degrees of freedom chosen for the fixed factor Y, as illustrated in Example 1.2. Based on this observation, we want to increase expressivity in the factor Z by increasing the non-negative range of Y. Corollary 3.2 shows the equivalence between subset relations of the non-negative ranges and non-negative matrix factorizations.

COROLLARY 3.2. *Let* $Y, V \in \mathbb{R}^{n_1 \times k}_{\geq 0}$. *Then* $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ *if and only if there exists an* $N \in \mathbb{R}^{k \times k}_{\geq 0}$ *such that* $Y = VN$.

Proof. Let $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$. Then, for each $j \in [k]$ there exists an $N_{:,j} \in \mathbb{R}^k_{\geq 0}$ with $Y_{:,j} = Ye_j = VN_{:,j}$, i.e., $Y = VN$ holds true. Let $Y = VN$ with $N \geq 0$. Then $Yz = V(Nz) \in \operatorname{range}_{\geq 0}(V)$ as $Nz \geq 0$ for all $z \geq 0$, i.e., $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ holds true. ∎

Extending [9, Lemma 18], the following subset relation between the (non-negative) ranges of Y and V holds true.

LEMMA 3.3. *Let* $Y, V \in \mathbb{R}^{n_1 \times k}_{\geq 0}$ *and* $N \in \mathbb{R}^{k \times k}_{\geq 0}$ *be regular such that* $Y = VN$. *Then* $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V) \subseteq \operatorname{range}(Y) \cap \mathbb{R}^{n_1}_{\geq 0}$ *and* $\operatorname{range}(Y) = \operatorname{range}(V)$ *hold true.*

Proof. First, $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ follows directly from Corollary 3.2. Similarly, for any $w \in \mathbb{R}^k_{\geq 0}$ we have $Vw = Y(N^{-1}w) \in \operatorname{range}(Y)$. The identity $\operatorname{range}(Y) = \operatorname{range}(V)$ follows directly from the regularity of N. ∎

Furthermore, we formulate a specific variant of the general vertex-representation theorem for polytopes [52, Theorem 2.15] for the non-negative range in Theorem 3.5.

THEOREM 3.4 (Vertex representation of polytopes [52]). *Any polytope* $P = \operatorname{conv}(S) := \{\sum_{v \in S} \lambda_v v \colon \lambda_v \geq 0, \; \sum_{v \in S} \lambda_v = 1\}$ *for a finite set* $S \subseteq \mathbb{R}^n$ *is the convex hull of its vertices* $T \subseteq P$, *that is,* $P = \operatorname{conv}(T)$.

THEOREM 3.5. *Let* $Y \in \mathbb{R}^{n_1 \times k}_{\geq 0}$ *be non-zero. Then the following statements hold true:*

(i) There exists a non-empty subset $\mathcal{I} \subseteq [k]$ *of columns with minimal number of elements such that* $\operatorname{range}_{\geq 0}(Y) = \operatorname{range}_{\geq 0}(Y_{:,\mathcal{I}})$.

(ii) All columns of $Y_{:,\mathcal{I}}$ *are extreme with respect to* $Y_{:,\mathcal{I}}$.

(iii) If no column of Y *is a multiple of another one, i.e.,* $Y_{:,i} \neq \lambda Y_{:,j}$ *for all* $\lambda \geq 0$ *and* $i \neq j$, *then* $\mathcal{I} \neq \emptyset$ *is uniquely defined as the index set of all extreme columns in* Y.

(iv) If Y *contains only extreme columns, then* $\mathcal{I} = [k]$ *is already minimal in the above sense.*

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Proof. The statement can be derived as a special case of the vertex representation theorem [52, Theorem 2.15] above. For the complete proof and explicit construction of the index set $\mathcal{I}$, we refer to Appendix C. ∎

Moreover, two matrices spanning the same non-negative range are equal up to a specific type of transformation matrix, called monomial, also known as generalized permutations:

DEFINITION 3.6 (Monomial matrix). *A regular matrix $M \in \mathbb{R}_{\geq 0}^{k\times k}$ is called a monomial matrix if and only if $M = PD$, where $P \in \{0,1\}^{k\times k}$ is a permutation and $D = \operatorname{diag}(m_1, \ldots, m_k)$ is a diagonal matrix with $m_\ell > 0$ for all $\ell \in [k]$. We denote the set of all $k \times k$ monomial matrices as*

$$\operatorname{Mono}_k := \{M \in \mathbb{R}_{\geq 0}^{k\times k} \colon M \text{ is monomial}\}.$$

Non-negative matrices with non-negative inverse are exactly the monomial ones; see, e.g., [30, Lemma 1.1].

LEMMA 3.7 ([30]). *A regular non-negative matrix $M \in \mathbb{R}_{\geq 0}^{k\times k}$ is inverse non-negative if and only if M is a monomial matrix.*

As motivated above, the non-negative range is invariant under monomial transformation.

LEMMA 3.8. *Let $Y \in \mathbb{R}_{\geq 0}^{n_1\times k}$ have only extreme columns. Then each non-zero element of its kernel, $x \in \operatorname{kernel}(Y) := \{x \in \mathbb{R}^k \colon Yx = \mathbf{0}_{n_1}\}$, $x \neq \mathbf{0}_k$, has at least two negative entries, i.e., there exist $j \neq \ell$ with $x_j, x_\ell < 0$.*

Proof. As $Y \geq 0$ has no zero column, each element $x \neq \mathbf{0}_k$ of its kernel must have at least one negative entry $x_\ell < 0$ for an $\ell \in [k]$. Assume that $x_i \geq 0$ for all $i \neq \ell$. Then due to $x \in \operatorname{kernel}(Y)$, we can write

$$Y_{:,\ell} = \sum_{i\neq\ell} \frac{x_i}{|x_\ell|} Y_{:,i},$$

which contradicts the assumption that $Y_{:,\ell} \notin \operatorname{range}_{\geq 0}(Y_{:,\neq\ell})$. ∎

LEMMA 3.9. *Let $Y \in \mathbb{R}_{\geq 0}^{n_1\times k}$ have only extreme columns and let $V \in \mathbb{R}_{\geq 0}^{n_1\times k}$. Then $\operatorname{range}_{\geq 0}(Y) = \operatorname{range}_{\geq 0}(V)$ holds true if and only if there exists a monomial matrix $M \in \operatorname{Mono}_k$ such that $Y = VM$.*

Proof. "⇒" Let $\operatorname{range}_{\geq 0}(Y) = \operatorname{range}_{\geq 0}(V)$. Using Corollary 3.2, there exists an $N \in \mathbb{R}_{\geq 0}^{k\times k}$ such that $Y = VN$ and similarly $M \in \mathbb{R}_{\geq 0}^{k\times k}$ with $V = YM$. Thus, we have

$$Y = VN = YMN \quad \Longleftrightarrow \quad Y(MN - \operatorname{Id}_k) = \mathbf{0}_{n_1\times k}.$$

For any fixed $\ell \in [k]$, we show that $x := (MN - \operatorname{Id}_k)_{:,\ell} \in \operatorname{kernel}(Y)$ has at most one negative entry, which by Lemma 3.8 implies that x is already zero. By construction and the fact that $M, N \geq 0$, directly $x_\ell \geq -1$ and $x_i \geq 0$ for all $i \neq \ell$ follow. Thus, $x \in \operatorname{kernel}(Y)$ has at most one negative entry, which implies that $x = \mathbf{0}_k$ must hold true. As $\ell \in [k]$ was chosen arbitrarily, it follows that $MN = \operatorname{Id}_k$. Due to $M, N \geq 0$, Lemma 3.7 implies that M and N are monomial matrices.

"⇐" Let $Y = VM$, then directly $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ holds true as $M \geq 0$ and $\operatorname{range}_{\geq 0}(V) \subseteq \operatorname{range}_{\geq 0}(Y)$ with $M^{-1} \geq 0$, i.e., $\operatorname{range}_{\geq 0}(Y) = \operatorname{range}_{\geq 0}(V)$ follows. ∎

THEOREM 3.10. *Let $Y \in \mathbb{R}_{\geq 0}^{n_1\times k}$ have only extreme columns and $V \in \mathbb{R}_{\geq 0}^{n_1\times k}$. Then $\operatorname{range}_{\geq 0}(Y) \subsetneq \operatorname{range}_{\geq 0}(V)$ holds true if and only if there exists a non-monomial $N \in \mathbb{R}_{\geq 0}^{k\times k}$ with $Y = VN$. In this case, N has only extreme columns.*

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Proof. "⇒" Let $\operatorname{range}_{\geq 0}(Y) \subsetneq \operatorname{range}_{\geq 0}(V)$. By Corollary 3.2, there exists an $N \in \mathbb{R}_{\geq 0}^{k\times k}$ with $Y = VN$ using Corollary 3.2. Using Lemma 3.9, N is non-monomial. Assume N has a non-extreme column $N_{:,\ell} = \sum_{j\neq\ell} \lambda_j N_{:,j} \geq 0$ with $\lambda_j \geq 0$. Then also $Y_{:,\ell} = VN_{:,\ell} = \sum_{j\neq\ell} \lambda_j VN_{:,j} = \sum_{j\neq\ell} \lambda_j Y_{:,j}$ would be non-extreme.

"⇐" Let $Y = VN$ for a non-monomial $N \in \mathbb{R}_{\geq 0}^{k\times k}$. Corollary 3.2 directly implies that $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ holds true. Assuming $\operatorname{range}_{\geq 0}(Y) = \operatorname{range}_{\geq 0}(V)$, Lemma 3.9 would imply that N is monomial, which contradicts the assumptions. ∎

In the context of (exact) non-negative matrix factorization $B = YZ^T$, Lemma 3.11 describes the relationship between maximizing the non-negative range of Y and minimizing the non-negative range of Z:

LEMMA 3.11. *Let $Y \in \mathbb{R}_{\geq 0}^{n_1\times k}$, $Z \in \mathbb{R}_{\geq 0}^{n_2\times k}$ with $\operatorname{rank}(Z) = k$, and $B = YZ^T$. Then there exists $V \in \mathbb{R}_{\geq 0}^{n_1\times k}$ with $\operatorname{range}_{\geq 0}(Y) \subsetneq \operatorname{range}_{\geq 0}(V)$ if and only if there exists $W \in \mathbb{R}^{n_2\times k}$ with $\operatorname{col}(B^T) \subseteq \operatorname{range}_{\geq 0}(W) \subsetneq \operatorname{range}_{\geq 0}(Z)$.*

Proof. "⇒" Assume there exists such V. Then due to $\operatorname{range}_{\geq 0}(Y) \subset \operatorname{range}_{\geq 0}(V)$, there is a non-monomial $N \in \mathbb{R}_{\geq 0}^{k\times k}$ with $Y = VN$ and hence $YZ^T = V(ZN^T)^T$ holds true. We define $W := ZN^T$. With $B^T = WV^T$ and $V \geq 0$, it follows that $\operatorname{col}(B^T) \subset \operatorname{range}_{\geq 0}(W)$. As N^T is non-monomial, we have shown the existence of W as required.

"⇐" Assume there exist a non-monomial $N \in \mathbb{R}_{\geq 0}^{k\times k}$ with $W = ZN^T$ and a non-negative $V \in \mathbb{R}_{\geq 0}^{n_1\times k}$ with $B^T = WV^T$. Let $Z^\dagger$ be the pseudo-inverse of Z^T. As $\operatorname{rank}(Z) = k$, we have $Z^TZ^\dagger = \operatorname{Id}_k$, and, by assumption, $Y = YZ^TZ^\dagger = BZ^\dagger = VW^TV = VNZ^TZ^\dagger = VN$ follows. As N is non-monomial, $\operatorname{range}_{\geq 0}(Y) \subsetneq \operatorname{range}_{\geq 0}(V)$ must hold true. ∎

REMARK 3.12 (On Lemma 3.11). Note that, in Lemma 3.11, the statements $\operatorname{col}(B) \subseteq \operatorname{range}_{\geq 0}(V)$ and $W \geq 0$ are each redundant, following from $\operatorname{col}(B) \subseteq \operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ and $\operatorname{col}(W) \subseteq \operatorname{range}_{\geq 0}(Z) \subseteq \mathbb{R}_{\geq 0}^{n_2}$.

Lemma 3.11 can also be applied to $Y = VN$ with regular N. Then, in the sense of Lemma 3.11, maximizing the non-negative range of V is equivalent to minimizing the non-negative range of N under the above constraints.

Further, Lemma 3.11 characterizes the (essential) uniqueness of the non-negative matrix factorization $B = YZ^T$ and thus relates, for example, to [25, Theorem 1] (for $\operatorname{rank}(B) = k$) and [19]. There, the uniqueness of non-negative matrix factorizations is studied based on the non-negative range and its dual space.

3.2. Geometric perspective on increase of non-negative range. Next, we want to illustrate our idea of increasing the non-negative range guided by the graphical representations in [9, Section 4.3] and recap the definition of the unit simplex.

DEFINITION 3.13 (Unit simplex). *Let $n \in \mathbb{N}$ with $n > 1$. The $(n-1)$-dimensional unit simplex is defined as $\Delta^{n-1} := \{y \in \mathbb{R}_{\geq 0}^n \colon \|y\|_1 = 1\}$.*

As the non-negative range is invariant under monomial transformation, see Lemma 3.9, we can assume that $\|Y_{:,j}\|_1 = 1$ holds true for all $j \in [k]$, i.e., $\operatorname{col}(Y) \subseteq \Delta^{n_1-1}$. Then we can illustrate the non-negative range of Y by plotting the intersection $\operatorname{range}_{\geq 0}(Y) \cap \Delta^{n_1-1}$ in Examples 3.14 and 3.15. Note that for $\operatorname{col}(Y) \subseteq \Delta^{n_1-1}$, this intersection is the set of all convex combinations of $\operatorname{col}(Y)$.

EXAMPLE 3.14 (Intersection $\operatorname{range}_{\geq 0}(Y) \cap \Delta^2$ for $Y \in \mathbb{R}_{\geq 0}^{3\times 2}$). Figure 3.1 illustrates the idea of increasing the non-negative range of $Y \in \mathbb{R}_{\geq 0}^{3\times 2}$: on the left, we have the two column vectors in black which span the non-negative range. All convex combinations, i.e., $\operatorname{range}_{\geq 0}(Y) \cap \Delta^2$, are highlighted in dark blue. As we are only interested in the intersection with the unit simplex, we can zoom in, as shown on the right, where the column vectors of Y correspond to points in Δ^2. If we now want to increase the non-negative range of Y and

ETNA
Kent State University and
Johann Radon Institute (RICAM)

still keep all convex combinations $\text{range}_{\geq 0}(Y) \cap \Delta^2$ included, the only possible way is to move the columns along the dashed line, which is $\text{range}(Y) \cap \Delta^2$, towards its boundary; see Lemma 3.3. This leads to $V \geq 0$ (highlighted in bordeaux) with maximal non-negative range such that $\text{range}_{\geq 0}(Y) \subseteq \text{range}_{\geq 0}(V)$ holds true.

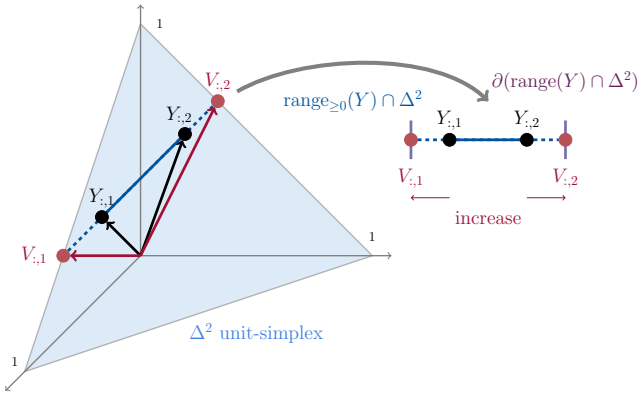


FIG. 3.1. *Illustration of uniquely maximizing the non-negative range of* $Y \in \mathbb{R}^{3\times 2}_{\geq 0}$: $\text{range}_{\geq 0}(Y) \subsetneq \text{range}_{\geq 0}(V)$.

Note that for $k \leq 2$ columns, the maximization of the non-negative range is quite trivial as the rank of Y is equal to its minimal non-negative factorization size. For $k > 2$, finding such a matrix V can be much more complicated. In particular, an optimal V might not be unique any more, as illustrated in Example 3.15.

EXAMPLE 3.15 (Intersection $\text{range}_{\geq 0}(Y) \cap \Delta^3$ for $Y \in \mathbb{R}^{4\times 3}_{\geq 0}$). Figures 3.2(a) and 3.2(b) show two different ways to increase the non-negative range of $Y \in \mathbb{R}^{4\times 3}_{\geq 0}$, where we modify Y column by column to increase its non-negative range. In Figure 3.2(a) we moved the first, then the second, and then the third column towards the boundary without losing points in $\text{range}_{\geq 0}(Y) \cap \Delta^3$. In Figure 3.2(b) we reverse the ordering of the columns, starting with the third, then the second, and lastly the first one. Depending on the ordering of moving the columns, we end up with two different matrices $V^{(a)}$ and $V^{(b)}$, which both increase the non-negative range. Both matrices $V^{(a)}$ and $V^{(b)}$ are optimal in the sense that their non-negative range cannot be increased further within $\text{range}(Y) \cap \Delta^3$ without losing other points from their respective non-negative range.

Note that extreme columns correspond to vertices and non-extreme columns to inner points of $\text{range}_{\geq 0}(Y) \cap \Delta^{n_1 - 1}$. This gives us further intuition for the existence of a minimal subset of generating columns in Theorem 3.5.

For all non-negative factorization $(V, N) \geq 0$ of $Y = VN$ with $\text{range}_{\geq 0}(Y) \subseteq \text{range}_{\geq 0}(V)$, in particular $\text{rank}(V) \geq \text{rank}(Y)$ must hold true, as otherwise there would be a column $Y_{:,j} \notin \text{range}(V)$, i.e., $Y_{:,j} \in \text{range}_{\geq 0}(Y) \setminus \text{range}_{\geq 0}(V)$, which would contradict the assumption. In [10, Theorem 3.6.], it is shown that finding a non-negative factorization $(V, N) \geq 0$ of Y with $\text{rank}(V) = \text{rank}(Y)$ is $\mathcal{NP}$-hard if $\text{rank}(Y) > 3$. This motivates us to focus on the following restriction allowing us to find a suitable non-negative factorization $(V, N) \geq 0$ in practice.

3.3. Restriction to inverse non-negative transfer matrices. Based on Corollary 3.2 and Theorem 3.10, one way to find a matrix $V \geq 0$ that increases the non-negative range of Y is to search for an inverse non-negative matrix $M \in \mathbb{R}^{k\times k}$ such that $V := YM \geq 0$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

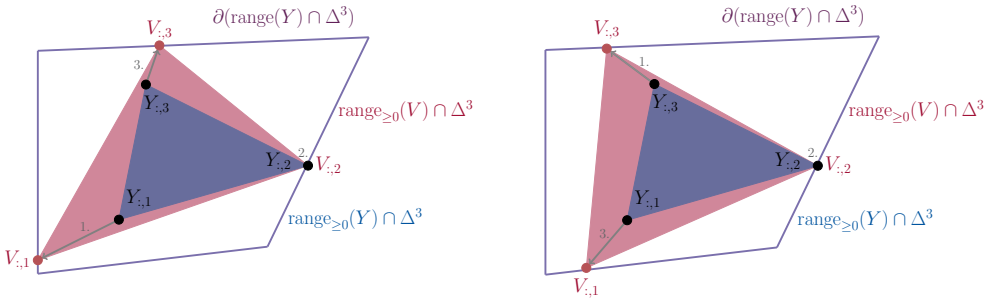


FIG. 3.2. *Illustration of increasing the non-negative range of* $Y \in \mathbb{R}^{4\times 3}_{\geq 0}$ *in two different ways:* $\mathrm{range}_{\geq 0}(Y) \subsetneq \mathrm{range}_{\geq 0}(V)$.

holds true. Then we have $Y = VN$ with $N := M^{-1} \geq 0$. However, optimizing over the set of inverse non-negative matrices seems to be a challenging task. Explicitly describing this semi-algebraic set for matrices of size $k \times k$ requires k^2 polynomial inequalities of degree less than or equal to k with up to $k!$ terms each [9]. Like [9], we therefore focus on a specific subset of inverse non-negative matrices denoted by $\mathcal{M}_Y$.

3.3.1. Construction of ansatz $\mathcal{M}_Y$. In this context, two important classes are Z- and M-matrices [3, Chapter 6]:

DEFINITION 3.16 (Z- and M-matrix). *A matrix $M \in \mathbb{R}^{k\times k}$ is called a Z-matrix if and only if all its off-diagonal entries are non-positive, i.e., $M_{i,j} \leq 0$ for all $i \neq j$. A Z-matrix M is called an M-matrix if and only if there exist a constant $s \geq 0$ and a non-negative matrix $C \in \mathbb{R}^{k\times k}_{\geq 0}$ such that $M = s\,\mathrm{Id}_k - C$ holds true, where $s \geq |\lambda|$ for all eigenvalues λ of C.*

Regular M-matrices can be characterized using the following equivalences; see, e.g., [3, Chapter 6].

LEMMA 3.17 (Inverse non-negative M-matrix). *For a Z-matrix $M \in \mathbb{R}^{k\times k}$, the following statements are equivalent:*

(i) M *is a regular M-matrix.*
(ii) M *is regular and inverse non-negative, i.e., $M^{-1} \geq 0$.*
(iii) There exists an $x \in \mathbb{R}^k_{>0}$ with $Mx > 0$.
(iv) M *has positive diagonal entries and there exists a diagonal matrix $D = \mathrm{diag}(m_1, \ldots, m_k)$ with $m_\ell > 0$ for all ℓ such that MD is strictly diagonal dominant.*

Proof. See [3, Chapter 6 $(N_{38})and(I_{28})$]. ∎

Similar to [9], we consider the following set of Z-matrices.

NOTATION 3.18. For $Y \in \mathbb{R}^{n_1\times k}_{\geq 0}$, we denote

$$\mathcal{M}_Y := \{M \in \mathbb{R}^{k\times k} \colon M_{\ell,\ell} > 0,\ M_{\ell,j} \leq 0\ \forall\, \ell \neq j,\ YM \geq 0\}.$$

REMARK 3.19 (Expressing $\mathcal{M}_Y$ via linear (in)equality constraints). As the non-negative range is invariant under monomial transformation (see Lemma 3.9), we can normalize all matrices $M \in \mathcal{M}_Y$ by fixing $M_{\ell,\ell} = 1$ for all $\ell \in [k]$ (rather than requiring $M_{\ell,\ell} > 0$). All matrices $M \in \mathcal{M}_Y$ with unit diagonal can be characterized by k linear equality constraints and $(n_1 + k - 1)k$ linear inequality constraints.

In the following, we derive conditions under which an element $M \in \mathcal{M}_Y$ is inverse non-negative. In this case, we can replace the current non-negative factorization $(Y, Z) \geq 0$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

by $(YM, ZM^{-T}) \geq 0$ without changing the represented matrix $X = \tau_k(Y, Z)$ or losing the non-negativity. To do so, we start with some conditions which are quite easy to check in practical applications.

3.3.2. Inverse non-negativity of $\mathcal{M}_Y$: practical perspective. One way to characterize inverse non-negative matrices in $\mathcal{M}_Y$, which is quite easy to prove, is given in Theorem 3.20.

THEOREM 3.20 (Inverse non-negativity of $\mathcal{M}_Y$). *Let $Y \in \mathbb{R}^{n_1 \times k}_{\geq 0}$ have no zero column and $M \in \mathcal{M}_Y$. Then either*

(i) $V := YM$ has at least one zero column, or

(ii) M is an inverse non-negative M-matrix, i.e., $M^{-1} \geq 0$.

Proof. Let $x := Y^T\mathbf{1}$. Then $x > 0$ holds true, as $Y \geq 0$ has no zero column. Further, we have $M^T x = M^T Y^T \mathbf{1} = V^T \mathbf{1} \geq 0$. Then either V has a zero column or $M^T x > 0$. As $M \in \mathcal{M}_Y$, M and M^T are Z-matrices. Thus, Lemma 3.17 implies that M^T is a regular inverse non-negative M-matrix, as $M^T x > 0$ holds true for $x > 0$, and thereby M. ∎

REMARK 3.21. Here, we do not need our transfer matrices to be regular M-matrices, but inverse non-negative. For this reason, one could also, for instance, use products $M = \prod_\mu M^{(\mu)}$ with regular $M^{(\nu)} \in \mathcal{M}_{Y \prod_{\mu=1}^{\nu-1} M^{(\mu)}}$ for all ν as $M^{-1} \geq 0$.

If $V := YM$ has zero columns, then, analogously to the orthogonalization of rank-deficient matrices, Theorem 3.20 allows us to delete redundant columns using inverse non-negative matrices (independent of the regularity of M itself):

LEMMA 3.22. *Let $Y \in \mathbb{R}^{n_1 \times k}_{\geq 0}$, $M \in \mathcal{M}_Y$, $V := YM$, and $V_{:,\ell} = \mathbf{0}$ for an $\ell \in [k]$. Then the following hold true:*

(i) $Y_{:,\ell} \in \operatorname{range}_{\geq 0}(Y_{:,\neq\ell})$, i.e., $\operatorname{range}_{\geq 0}(Y_{:,\neq\ell}) = \operatorname{range}_{\geq 0}(Y)$.

(ii) $\widetilde{M} := \mathrm{Id}_k + (M_{:,\ell} - e_\ell) \otimes e_\ell$ is an inverse non-negative M-matrix with $\widetilde{M}^{-1} = \mathrm{Id}_k - (M_{\ell,\ell})^{-1}(M_{:,\ell} - e_\ell) \otimes e_\ell$ and $Y\widetilde{M} = \begin{bmatrix} Y_{:,<\ell} & \mathbf{0} & Y_{:,>\ell} \end{bmatrix}$.

Proof.

(i) Due to $M_{\ell,\ell} > 0 \geq M_{j,\ell}$ for all $j \neq \ell$, we have $\mathbf{0}_{n_1} = V_{:,\ell} = YM_{:,\ell}$, which is equivalent to $Y_{:,\ell} = M_{\ell,\ell}^{-1} \sum_{j \neq \ell} |M_{j,\ell}| Y_{:,j} \in \operatorname{range}_{\geq 0}(Y_{:,\neq\ell})$. Thus, $\operatorname{range}_{\geq 0}(Y_{:,\neq\ell}) = \operatorname{range}_{\geq 0}(Y)$ follows.

(ii) $\widetilde{M}$ is a Z-matrix and for $x := \mathbf{1} - \sum_{j \neq \ell} M_{j,\ell}\, e_\ell > 0$, we have $\widetilde{M}^T x = \mathbf{1} + e_\ell > 0$. Due to Lemma 3.17, $\widetilde{M}$ is an inverse non-negative M-matrix. The formulas for $Y\widetilde{M}$ and $\widetilde{M}^{-1}$ follow by simple calculations. ∎

Note that Lemma 3.22(i) together with Theorem 3.20 in particular implies that, if Y has only extreme columns, then all $M \in \mathcal{M}_Y$ are inverse non-negative.

REMARK 3.23 (On Lemma 3.22(ii)). On the one hand, Lemma 3.22(ii) allows us to update $V := Y\widetilde{M} \geq 0$ and $W := Z\widetilde{M}^{-T} \geq 0$ such that $YZ^T = VW^T$ holds true. On the other hand, as the update V contains at least one zero column $V_{:,\ell} = \mathbf{0}$, we can reduce the factorization size to $k - 1$ and repeat the quasi-orthogonalization for $\widetilde{Y} := V_{:,\neq\ell}$ and $\widetilde{Z} := W_{:,\neq\ell}$. Alternatively, we can replace this zero column $V_{:,\ell}$ by a random one, set the corresponding column in W to zero, and repeat the quasi-orthogonalization for $\widetilde{Y} := V$ and $\widetilde{Z} := W$ without reducing the factorization size k. Consequently, we can repeat one of these strategies until we come to the case (ii) of Theorem 3.20.

Note that, even if $V := YM$ contains at least one zero column (see Theorem 3.20(i)), $M \in \mathcal{M}_Y$ can still be inverse non-negative. We will theoretically analyze this further in the next section, using additional characterizations of the regularity of M-matrices. In practice, however, it may be easier to proceed as described in Remark 3.23.

3.3.3. Inverse non-negativity of $\mathcal{M}_Y$: theoretical perspective. Besides the equivalences in Lemma 3.17, inverse non-negativity of a Z-matrix can also be concluded if it is

ETNA
Kent State University and
Johann Radon Institute (RICAM)

irreducible diagonal dominant; see Definition 3.24 and Lemma 3.25 [42, Theorem 3]. We will use this result to prove inverse non-negativity of $M \in \mathcal{M}_Y$ in Theorem 3.29. This idea is similar to [9, Theorem 21] but in a more general way to allow for its application to $M \in \mathcal{M}_Y$.

DEFINITION 3.24 (Irreducible and diagonal dominant matrix). *A matrix $M \in \mathbb{R}^{k\times k}$ is called* irreducible *if and only if there does not exist a permutation $P \in \{0,1\}^{k\times k}$ such that*

$$P^T M P = \begin{bmatrix} M^{(1,1)} & M^{(1,2)} \\ \mathbf{0} & M^{(2,2)} \end{bmatrix} \quad \text{with} \quad M^{(i,i)} \in \mathbb{R}^{k_i \times k_i},\ k = k_1 + k_2,$$

holds true. Matrix M is called diagonal dominant *if and only if*

$$M_{\ell,\ell} \geq \sum_{j\neq\ell} |M_{j,\ell}| \tag{3.1}$$

holds true for all $\ell \in [k]$, and it is called strictly diagonal dominant *if and only if all inequalities in (3.1) hold strictly. Finally, matrix M is called* irreducible diagonal dominant *if and only if M is irreducible, diagonal dominant, and (3.1) holds strictly for at least one $\ell \in [k]$.*

LEMMA 3.25 ([42]). *If a Z-matrix $M \in \mathbb{R}^{k\times k}$ is irreducible diagonal dominant, then M is an inverse non-negative M-matrix.*

To use Lemma 3.25 for $M \in \mathcal{M}_Y$, we replace the diagonal dominance in Lemma 3.26.

LEMMA 3.26. *If a Z-matrix $M \in \mathbb{R}^{k\times k}$ is irreducible and $\mathbf{0}_k \neq \mathbf{1}_k^T M \geq 0$ holds true, then M is an inverse non-negative M-matrix.*

Proof. We prove that the condition $\mathbf{0}_k \neq \mathbf{1}_k^T M \geq 0$ implies that M is diagonal dominant and (3.1) holds strictly for at least one $\ell \in [k]$. Then Lemma 3.25 implies that M is an inverse non-negative M-matrix. For $\ell \in [k]$, it holds true that $0 \leq \mathbf{1}_k^T M_{:,\ell} = M_{\ell,\ell} + \sum_{j\neq\ell} M_{j,\ell}$, which is equivalent to $M_{\ell,\ell} \geq \sum_{j\neq\ell} |M_{j,\ell}|$, as $M_{j,\ell} \leq 0$ for all $j \neq \ell$. As $\mathbf{1}_k^T M \neq \mathbf{0}_k$ holds true, there exists at least one $i \in [k]$ such that $\mathbf{1}_k^T M_{:,i} > 0$, which is equivalent to $M_{i,i} > \sum_{j\neq i} |M_{j,i}|$. ∎

We prove that each $M \in \mathcal{M}_Y$ fulfills the second condition of Lemma 3.26 if Y does not contain multiple columns.

LEMMA 3.27. *Let $Y \in \mathbb{R}_{\geq 0}^{n_1\times k}$ with $\|Y_{:,j}\|_1 = 1$ for all $j \in [k]$. Then each $M \in \mathcal{M}_Y$ fulfills $\mathbf{1}_k M \geq 0$.*

Proof. Since $YM \geq 0$ holds true, we have $\mathbf{1}_k^T M = (\mathbf{1}_{n_1}^T Y)M = \mathbf{1}_{n_1}^T (YM) \geq 0$. ∎

COROLLARY 3.28. *Let $Y \in \mathbb{R}_{\geq 0}^{n_1\times k}$ have no column which is a multiple of another one. Then, for each Z-matrix M with $M_{\ell,\ell} > 0$ for all $\ell \in [k]$, the product $YM \neq \mathbf{0}_{n_1\times k}$.*

Proof. Using Theorem 3.5 there exists at least one extreme column $Y_{:,j} \notin \operatorname{range}_{\geq 0}(Y_{:,\neq j})$. Assume $YM_{:,j} = \mathbf{0}_{n_1}$. Then, due to $M_{j,j} > 0$, it holds true that $\mathbf{0}_{n_1} = YM_{:,j} = M_{j,j}Y_{:,j} \sum_{\ell\neq j} M_{\ell,j}Y_{:,\ell}$. This is equivalent to $Y_{:,j} = \sum_{\ell\neq j} |M_{\ell,j}|/M_{j,j}Y_{:,\ell}$, i.e., $Y_{:,j} \in \operatorname{range}_{\geq 0}(Y_{:,\neq j})$, which contradicts the assumption. ∎

Combining Lemmas 3.26 and 3.27, we conclude the following result on the inverse non-negativity of $M \in \mathcal{M}_Y$.

THEOREM 3.29 (Inverse non-negativity of $\mathcal{M}_Y$). *Let $Y \in \mathbb{R}_{\geq 0}^{n_1\times k}$ with $\|Y_{:,j}\|_1 = 1$ for all $j \in [k]$ and $Y_{:,j} \neq Y_{:,\ell}$ for all $j \neq \ell$. Then each $M \in \mathcal{M}_Y$ is an inverse non-negative M-matrix.*

Proof. As $M \in \mathcal{M}_Y$ is a Z-matrix with positive diagonal, Lemma 3.27 and Corollary 3.28 imply that $\mathbf{0}_k^T \neq \mathbf{1}_k^T M \geq 0$. If M is irreducible, then the statement follows directly using Lemma 3.26. Otherwise, there exists a permutation $P \in \{0,1\}^{k\times k}$ such that

$$P^T M P = \begin{bmatrix} M^{(1,1)} & M^{(1,2)} \\ \mathbf{0} & M^{(2,2)} \end{bmatrix}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Let $\mathcal{J}_\ell \subseteq [k]$ denote the corresponding index set such that $M_{\mathcal{J}_\ell,\mathcal{J}_\ell} = M^{(\ell,\ell)} \in \mathbb{R}^{k_\ell \times k_\ell}$. Hence, the off-diagonal block $M^{(1,2)} = M_{\mathcal{J}_1,\mathcal{J}_2} \leq 0$ and the diagonal blocks $M^{(\ell,\ell)}$ are again Z-matrices with positive diagonal and

$$Y_{:,\mathcal{J}_\ell} M^{(\ell,\ell)} = Y_{:,\mathcal{J}_\ell} M_{\mathcal{J}_\ell,\mathcal{J}_\ell} \geq -Y_{:,\neq\mathcal{J}_\ell} M_{\neq\mathcal{J}_\ell,\mathcal{J}_\ell} \geq 0,$$

i.e., $M^{(\ell,\ell)} \in \mathcal{M}_{Y_{:,\mathcal{J}_\ell}}$ follows. Thus, if $M^{(\ell,\ell)}$ is irreducible, then Lemma 3.26 implies that $(M^{(\ell,\ell)})^{-1} \geq 0$ holds true. Otherwise, we can also permute $M^{(\ell,\ell)}$ into a similar block structure and argue analogously until all diagonal blocks are irreducible and thus inverse non-negative. Using the following block inversion recursively for inverse non-negative matrices A and C and a non-positive $B \leq 0$,

$$\begin{bmatrix} A & B \\ \mathbf{0} & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & -A^{-1}BC^{-1} \\ \mathbf{0} & C^{-1} \end{bmatrix} \geq 0,$$

we also conclude that M is an inverse non-negative M-matrix. ∎

Using Corollary 3.28, Theorem 3.29 can easily be extended to matrices Y where no column is a multiple of another one.

COROLLARY 3.30. *Let $Y \in \mathbb{R}_{\geq 0}^{n_1 \times k}$ have no column which is a multiple of another one, i.e., $Y_{:,j} \neq \lambda Y_{:,\ell}$ holds true for all $\lambda \geq 0$ and $j \neq \ell$. Then each $M \in \mathcal{M}_Y$ is an inverse non-negative M-matrix.*

Proof. As Y has no zero column, $D := \operatorname{diag}(\|Y_{:,1}\|_1, \ldots, \|Y_{:,k}\|_1)$ is inverse non-negative. Then $\widetilde{Y} := YD^{-1} \geq 0$ fulfills $\|\widetilde{Y}_{:,j}\|_1 = 1$ and $\widetilde{Y}_{:,j} \neq \widetilde{Y}_{:,\ell}$ for all $j \neq \ell$. Hence, Theorem 3.29 implies that each $\widetilde{M} \in \mathcal{M}_{\widetilde{Y}}$ is inverse non-negative. Further, each $M \in \mathcal{M}_Y$ can be written as $M = D\widetilde{M}$ for an $\widetilde{M} \in \mathcal{M}_{\widetilde{Y}}$ and is therefore also inverse non-negative. ∎

Theorems 3.29 and 3.20 cover different perspectives: Theorem 3.29 shows the regularity of $\mathcal{M}_Y$ for Y with distinct columns, i.e., focusing on Y, and Theorem 3.20 shows the regularity of a given M if $V := YM$ has no zero column, i.e., focusing on V. However, both indicate that an $M \in \mathcal{M}_Y$ can only be singular if Y contains a non-extreme column, which can be removed (or replaced) without decreasing its non-negative range; see Remark 3.23. Here, we are interested in finding not only an inverse non-negative matrix $M \in \mathcal{M}_Y$ but one that allows us to improve the expressivity in Z by increasing the non-negative range of $V := YM$.

3.3.4. Increasing the non-negative range via $\mathcal{M}_Y$. We prove in Theorem 3.31 that, for any non-diagonal $M \in \mathcal{M}_Y$ and $V := YM$ having no zero column, the corresponding update V increases the non-negative range compared to Y.

THEOREM 3.31 (Increasing the non-negative range over $\mathcal{M}_Y$). *Let $p \geq 1$, $Y \in \mathbb{R}_{\geq 0}^{n_1 \times k}$, $M \in \mathcal{M}_Y$ be inverse non-negative, and $V := YM$ with $\|Y_{:,j}\|_p = \|V_{:,j}\|_p = 1$ for all $j \in [k]$. Then the following statements hold true:*

(i) $0 \leq M^{-1} \leq 1$ *and* $\|M_{:,\ell}^{-1}\|_p \leq 1$ *for all* $\ell \in [k]$.

(ii) If $p = 1$, then also $\|M_{:,\ell}^{-1}\|_1 = 1$ for all $\ell \in [k]$, i.e., each column of Y is a convex combination of the columns of V.

(iii) $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$.

(iv) If V has only extreme columns, then either $M = \operatorname{Id}_k$ or $\operatorname{range}_{\geq 0}(Y) \subsetneq \operatorname{range}_{\geq 0}(V)$.

Proof.

(i) As $Y, V \geq 0$ with $\|Y_{:,j}\|_p = \|V_{:,j}\|_p = 1$ for all $j \in [k]$, and $M^{-1} \geq 0$ holds true, using Lemma B.2 we have $1 = \|Y_{:,\ell}\|_p = \|VM_{:,\ell}^{-1}\|_p \geq \|M_{:,\ell}^{-1}\|_p$ for all $\ell \in [k]$. Thus, it follows that $M^{-1} \leq 1$.

(ii) For $p = 1$, the equality follows directly from above with Lemma B.2.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

(iii) We have $Yz = VM^{-1}z$ with $M^{-1}z \geq 0$ for all $z \in \mathbb{R}^k_{\geq 0}$, i.e., $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$.
(iv) If V has only extreme columns, then Lemma 3.9 implies that $\operatorname{range}_{\geq 0}(Y) = \operatorname{range}_{\geq 0}(VM)$ holds true if and only if $M \in \operatorname{Mono}_k$. Due to $M_{j,\ell} > 0 \Leftrightarrow j = \ell$ and $\|Y_{:,j}\|_p = \|V_{:,j}\|_p = 1$ for all j, this is equivalent to $M = \operatorname{Id}_k$. ∎

Note that Theorem 3.31 requires only non-zero columns in Y, via $\|Y_{:,j}\|_p = 1$ for all j, not extreme columns, which differs from the general formula in Theorem 3.10.

Combining Remark 3.23 and Theorem 3.31, there exists a finite sequence of updates using inverse non-negative matrices and restrictions which increases the non-negative range or keeps it from decreasing, while removing or replacing non-extreme columns.

As already illustrated in Example 3.15, the subset relation $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ defines only a partial ordering over the non-negative matrices of size $n_1 \times k$ with extreme columns modulo the monomial matrices Mono_k. For this reason, there can be multiple matrices with maximal non-negative range in the sense that their non-negative range cannot be further increased without losing other elements or their non-negativity. For example, the resulting matrices $V^{(a)}$ and $V^{(b)}$ in Figures 3.2(a) and 3.2(b) both have maximal non-negative range in this sense. But as neither $\operatorname{range}_{\geq 0}(V^{(a)}) \subseteq \operatorname{range}_{\geq 0}(V^{(b)})$ nor $\operatorname{range}_{\geq 0}(V^{(b)}) \subseteq \operatorname{range}_{\geq 0}(V^{(a)})$ holds true, there is no canonical way which one to choose.

Instead, we use a suitable measure $\mu_{\operatorname{range}_{\geq 0}}(\cdot)$ that defines a strict ordering and fulfills $\mu_{\operatorname{range}_{\geq 0}}(Y) \leq \mu_{\operatorname{range}_{\geq 0}}(YM)$ if and only if $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(YM)$, as well as $\mu_{\operatorname{range}_{\geq 0}}(Y) < \mu_{\operatorname{range}_{\geq 0}}(YM)$ if and only if $\operatorname{range}_{\geq 0}(Y) \subsetneq \operatorname{range}_{\geq 0}(YM)$ for all $M \in \mathcal{M}_Y$. However, it is not clear which measure is the more suitable. We propose a similar approach as shown in Examples 3.14 and 3.15 in the following section.

3.4. Column-wise increasing the non-negative range. As shown in Examples 3.14 and 3.15, our idea is to successively increase the non-negative range of Y by transforming Y column by column. In Theorem 3.32 we prove that replacing one extreme column in Y leads to a rise in its non-negative range if and only if there exists a transfer matrix of a specific structure (3.2) in $\mathcal{M}_Y$.

THEOREM 3.32. *Let $Y, V \in \mathbb{R}^{n_1 \times k}_{\geq 0}$ and let Y have no zero columns. For a fixed extreme column $Y_{:,\ell} \notin \operatorname{range}_{\geq 0}(Y_{:,\neq \ell})$ with $\ell \in [k]$, let $V_{:,\neq \ell} = Y_{:,\neq \ell}$ be given. Then $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ holds true if and only if there exists a vector $\beta \in \mathbb{R}^k$ with $\beta_\ell > 0$ and $\beta_j \leq 0$ for all $j \neq \ell$ such that $V = YM_\beta$, where*

$$M_\beta := \operatorname{Id}_k + (\beta - e_\ell) \otimes e_\ell \in \mathcal{M}_Y \quad \text{with} \quad M_\beta^{-1} = \operatorname{Id}_k - \beta_\ell^{-1}(\beta - e_\ell) \otimes e_\ell \geq 0. \tag{3.2}$$

In this case, $V_{:,\ell}$ is also extreme. Further, $\operatorname{range}_{\geq 0}(Y) \subsetneq \operatorname{range}_{\geq 0}(V)$ holds true if and only if there exists at least one $j \neq \ell$ with $\beta_j < 0$.

Proof. "⇒" Let $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ be fulfilled. Then there exists $N \in \mathbb{R}^{k \times k}_{\geq 0}$ such that $Y = VN$ holds true and, due to $Y_{:,\neq \ell} = V_{:,\neq \ell}$, we have $N_{:,j} = e_j$ for all $j \neq \ell$. We define $\alpha := N_{:,\ell} \geq 0$. Assuming $\alpha_\ell = 0$ would result in $Y_{:,\ell} = V\alpha = \sum_{j \neq \ell} \alpha_j Y_{:,j} \in \operatorname{range}_{\geq 0}(Y_{:,\neq \ell})$, which contradicts the assumption. For this reason, N is regular with $\det(N) = \alpha_\ell \det(\operatorname{Id}_{k-1}) = \alpha_\ell > 0$. We define $\beta_j := -\alpha_j/\alpha_\ell \leq 0$ and $\beta_\ell := 1/\alpha_\ell > 0$. Then we can write $M_\beta^{-1} := N \geq 0$ and $M_\beta := N^{-1}$ as above. The fact that $M_\beta \in \mathcal{M}_Y$ is inverse non-negative follows directly from (3.2). Assuming that $V_{:,\ell} \in \operatorname{range}_{\geq 0}(V_{:,\neq \ell})$, there exists a $\lambda \in \mathbb{R}^k_{\geq 0}$ such that $V_{:,\ell} = \sum_{j \neq \ell} \lambda_j V_{:,j}$ due to $\alpha \geq 0$ and $Y_{:,\neq \ell} = V_{:,\neq \ell}$. But then

$$Y_{:,\ell} = VN_{:,\ell} = \alpha_\ell V_{:,\ell} + \sum_{j \neq \ell} \alpha_j V_{:,j} = \sum_{j \neq \ell} (\alpha_\ell \lambda_j + \alpha_j) Y_{:,j} \in \operatorname{range}_{\geq 0}(Y_{:,\neq \ell})$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

contradicts the assumption, i.e., $V_{:,\ell}$ is an extreme column. Using Lemma 3.9, $\operatorname{range}_{\geq 0}(Y) \subsetneq \operatorname{range}_{\geq 0}(V)$ implies $\beta_j < 0$ for at least one $j \neq \ell$, as otherwise M_β would be monomial.

"$\Leftarrow$" Let $V = YM_\beta$ be given. As $N_\beta := M_\beta^{-1} \geq 0$ holds true, Corollary 3.2 directly implies that $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(V)$ is fulfilled. Let further $\beta_j < 0$ for at least one $j \neq \ell$. Assuming that $\operatorname{range}_{\geq 0}(Y) = \operatorname{range}_{\geq 0}(V)$ holds true, this then results in a contradiction similar to the argumentation in Lemma 3.9 using the facts that $Y_{:,\ell}$ is extreme and $Y_{:,\neq\ell} = V_{:,\neq\ell}$ holds true. ∎

Note that we already used matrices of the structure (3.2) in Lemma 3.22(ii) to set non-extreme columns in Y to zero.

An increase of the non-negative range by changing the extreme column $Y_{:,\ell}$ via Theorem 3.32, i.e., modifying Y by M_β with $\beta_j < 0$ for a $j \neq \ell$, is only possible if the support of $Y_{:,j}$ is a subset of the support of $Y_{:,\ell}$:

COROLLARY 3.33. *In the setting of Theorem 3.32: $\beta_j < 0$ for a $j \neq \ell$ is only possible if $Y_{i,\ell} = 0$ implies that $Y_{i,j} = 0$ for each $i \in [n_1]$.*

Proof. Assume that $\beta_j < 0$ holds true, but $Y_{i,\ell} = 0 < Y_{i,j}$ is satisfied. Then, due to $Y \geq 0$ and $\beta_{\neq\ell} \leq 0$, it would hold true that

$$0 \leq V_{i,\ell} = Y_{i,:}\beta = \underbrace{Y_{i,\ell}\beta_\ell}_{=0} + Y_{i,j}\beta_j + \sum_{\nu \neq i,\ell} \underbrace{Y_{i,\nu}\beta_\nu}_{\leq 0} \leq Y_{i,j}\beta_j < 0. \qquad \square$$

In [10, p. 219], it is explained that, due to the non-negativity constraints, the columns $Y_{:,j}$ typically share only a few non-zero entries (and analogously for Z). Assuming that the non-zero entries in the columns $Y_{:,j}$ were randomly distributed, Corollary 3.33 would indicate that finding $\beta_j < 0$, and thus increasing the non-negative range by M_β, becomes more difficult with increasing n_1.

Further, if all columns of Y and V in Theorem 3.32 have ℓ_1-norm of 1, then β_ℓ is determined by $\beta_{\neq\ell}$:

COROLLARY 3.34. *In the setting of Theorem 3.32: let $V = YM_\beta$ with $\|Y_{:,j}\|_1 = \|V_{:,j}\|_1 = 1$ for all $j \in [k]$ hold true. Then $\beta_\ell = 1 - \sum_{j\neq\ell} \beta_j$.*

Proof. We define $\alpha := (M_\beta^{-1})_{:,\ell} \geq 0$. Then it holds true that $1 = \|Y_{:,\ell}\|_1 = \|V\alpha\|_1 = \sum_j \alpha_j \|Y_{:,\ell}\|_1 = \sum_j \alpha_j = 1/\beta_\ell - \sum_{j\neq\ell} \beta_j/\beta_\ell \;\Leftrightarrow\; \beta_\ell = 1 - \sum_{j\neq\ell} \beta_j$. ∎

REMARK 3.35. For Y with $\|Y_{:,j}\|_1 = 1$ for all j, we can write

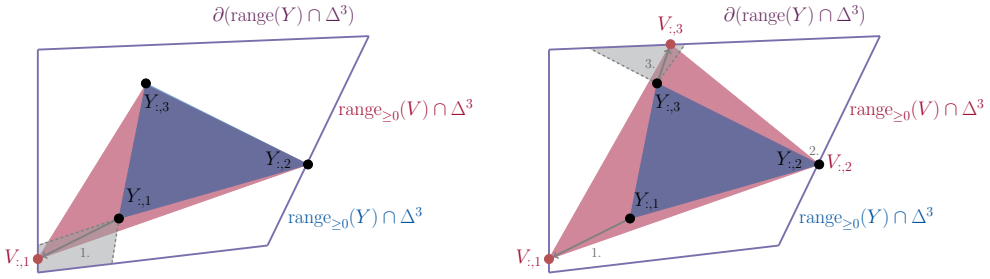
$$V_{:,\ell} = Y\beta = Y_{:,\ell} + \sum_{j\neq\ell} \beta_j (Y_{:,j} - Y_{:,\ell}) \in \Delta^{n_1-1}$$

using Corollary 3.34. From a geometrical point of view, this means that we move $Y_{:,\ell}$ in the opposite direction from each other column $Y_{:,j}$ for $j \neq \ell$. Figure 3.3 illustrates this for Y and $V \in \mathbb{R}^{4\times 3}_{\geq 0}$ from Figure 3.2(a). The possible locations to move $Y_{:,\ell}$ using an $M_\beta \in \mathcal{M}_Y$ are each highlighted in gray. As $Y_{:,2} \in \partial(\operatorname{range}(Y) \cap \Delta^3)$ holds true, it cannot be moved anywhere else.

Theorem 3.32 says that increasing the non-negative range of Y by changing only one single extreme column is equivalent to finding an $M_\beta \in \mathcal{M}_Y$ as in (3.2) with $\beta_{\neq\ell} \neq \mathbf{0}_{k-1}$. Assuming $\operatorname{col}(Y) \subseteq \Delta^{n_1-1}$, $Y_{:,\ell}$ to be an extreme column, and choosing $V_{:,\ell} \in \Delta^{n_1-1}$, one simple target function to find a suitable β is

$$\beta^+ \in \operatorname*{argmin}_{\beta\in\mathbb{R}^k} \sum_{j\neq\ell} \beta_j \quad \text{s.t. } \beta_j \leq 0 \,\forall\, j \neq \ell,\ Y\beta \in \Delta^{n_1-1}. \tag{3.3}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)



(a) (b)

FIG. 3.3. *Illustration to increase* $\mathrm{range}_{\geq 0}(Y)$ *column-wise using an* $M_\beta \in \mathcal{M}_Y$ *for* $Y \in \mathbb{R}^{4\times 3}_{\geq 0}$ *from Figure 3.2(a).*

As illustrated in Example 3.36, replacing $Y_{:,\ell}$ by $Y\beta^+$ before updating the next column may allow for further increase in the non-negative range. This strategy can be repeated until the columns in Y do not change any more.

EXAMPLE 3.36. Let $Y \in \mathbb{R}^{4\times 3}_{\geq 0}$ be defined as

$$Y := \begin{bmatrix} 0 & 1/3 & 0 \\ 1/3 & 1/3 & 0 \\ 1/3 & 0 & 1 \\ 1/3 & 1/3 & 0 \end{bmatrix}.$$

Then solving (3.3) independently, we obtain

$$\widetilde{\beta}^{(1)} := \begin{bmatrix} 3/2 \\ 0 \\ -1/2 \end{bmatrix}, \quad \widetilde{\beta}^{(2)} := \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \widetilde{\beta}^{(3)} := \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad \widetilde{V} = \begin{bmatrix} 0 & 1/3 & 0 \\ 1/2 & 1/3 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/3 & 0 \end{bmatrix},$$

with $\widetilde{V} := YM_{\widetilde{\beta}^{(1)}}$ for $M_{\widetilde{\beta}^{(1)}}$ as in (3.2). Then $\mathrm{range}_{\geq 0}(Y) \subsetneq \mathrm{range}_{\geq 0}(\widetilde{V})$ holds true since

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \in \mathrm{range}_{\geq 0}(\widetilde{V}) \setminus \mathrm{range}_{\geq 0}(Y).$$

Whereas solving (3.3) column by column and inserting each update leads to

$$\beta^{(1)} := \begin{bmatrix} 3/2 \\ 0 \\ -1/2 \end{bmatrix}, \quad \beta^{(2)} := \begin{bmatrix} -2 \\ 3 \\ 0 \end{bmatrix}, \quad \beta^{(3)} := \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad V = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 0 \\ 0 & 0 & 1 \\ 1/2 & 0 & 0 \end{bmatrix},$$

with $V := YM_{\beta^{(1)}}M_{\beta^{(2)}}M_{\beta^{(3)}}$ and $\mathrm{range}_{\geq 0}(Y) \subsetneq \mathrm{range}_{\geq 0}(\widetilde{V}) \subsetneq \mathrm{range}_{\geq 0}(V)$, as $e_1 \in \mathrm{range}_{\geq 0}(V) \setminus \mathrm{range}_{\geq 0}(\widetilde{V})$, i.e., V is preferable compared to $\widetilde{V}$ for an update of the second factor $Z \in \mathbb{R}^{n_2\times 3}_{\geq 0}$.

We summarize this procedure in Algorithm 3.1. Note that, instead of running the columns from $\ell = 1$ to $\ell = k$, one can instead choose any other order of columns.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 3.1: `quasi-ortho` for $\tau_k(Y, Z)$ w.r.t. Z

Input: $(Y, Z) \in \mathbb{R}^{n_1 \times k}_{\geq 0} \times \mathbb{R}^{n_2 \times k}_{\geq 0}$ (without zero columns)
Output: Updates $(Y^+, Z^+) \geq 0$ such that $\operatorname{range}_{\geq 0}(Y) \subseteq \operatorname{range}_{\geq 0}(Y^+)$ with $\|Y^+_{:,\ell}\|_1 = 1 \; \forall \ell$ and $Y^+(Z^+)^T = YZ^T$

1 Initialize $k = |\operatorname{col}(Y)|$
2 Compute $D := \operatorname{diag}(\|Y_{:,\ell}\|_1 \colon \ell = 1, \ldots, k)$
3 Update $Y \leftarrow YD^{-1}$ and $Z \leftarrow ZD^T$ // Column-wise rescaling
4 Initialize $M := \mathbf{0}_{k \times k}$
5 **while** $M \neq \mathrm{Id}_k$ **do**
6 Reset $M := \mathrm{Id}_k$
7 **for** $\ell = 1, \ldots, k$ **do**
8 Initialize $N := \mathrm{Id}_k$ // $N := M^{-1}$
9 Solve (3.3) for β^+
10 Update $M_{:,\ell} \leftarrow \beta^+$
11 Update $N_{:,\ell} \leftarrow -\beta^+/\beta^+_\ell$ and $N_{\ell,\ell} \leftarrow 1/\beta^+_\ell$
12 Update $Y_{:,\ell} \leftarrow Y\beta^+$
13 Update $Z \leftarrow ZN^T$
14 **if** $\|Y_{:,\ell}\|_1 = 0$ **then** // Replace zero column by random one
15 Choose $v := \operatorname{rand}(n_1, 1)$
16 Update $Y_{:,\ell} \leftarrow \dfrac{v}{\|v\|_1}$ and $Z_{:,\ell} \leftarrow \mathbf{0}_{n_2}$
17 **end**
18 **end**
19 **end**
20 **return** $(Y^+, Z^+) := (Y, Z)$

REMARK 3.37 (Relaxing (3.3)). In practical application, one could think of further relaxing the non-negativity constraint $Y\beta \geq 0$ in (3.3) by $Y\beta \geq -\varepsilon$ for a small $\varepsilon \geq 0$. This means solving

$$\beta^+ \in \operatorname*{argmin}_{\beta \in \mathbb{R}^k} \sum_{j \neq \ell} \beta_j \quad \text{s.t. } \beta_j \leq 0 \; \forall j \neq \ell, \; \beta_\ell = -\sum_{j \neq \ell} \beta_j, \; Y\beta \geq -\varepsilon \tag{3.4}$$

instead of (3.3) in line 9 and then project $Y \leftarrow \max\{YM, 0\}$ and $Z \leftarrow \max\{ZN^T, 0\}$ back to the non-negative orthant. In Section 4.2.7, such a relaxation may allow for slight improvements using $\varepsilon \in [10^{-16}, 10^{-8}]$, although the differences are quite small.

REMARK 3.38 (Generalization to TT tensors). The quasi-orthogonalization strategy in Algorithm 3.1 is easily extended to higher-dimensional tensors. We can perform the quasi-orthogonalization for a TT tensor $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu \in [d]})$ with respect to $\mathbf{X}_\nu$ similarly to the explanation in Algorithm 2.1 and Figure 2.4. We can use Algorithm 3.1 for the unfoldings $Y := \mathbf{X}_\mu{}^{(\{1,2\})}$ and $Z := \mathbf{X}_{\mu+1}{}^{(\{1\})}$ with $\mu = 1, \ldots, \nu - 1$, as well as for $Y := (\mathbf{X}_\eta{}^{(\{1\})})^T$ and $Z := (\mathbf{X}_{\eta-1}{}^{(\{1,2\})})^T$ with $\eta = d, \ldots, \nu + 1$.

In more detail, in the context of the three-dimensional example in Section 1.2, applying Algorithm 3.1 to $\mathbf{X} = \tau_k(\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3) \in \mathbb{R}^{n \times n \times n}_{\geq 0}$ corresponds to the following: In order to quasi-orthogonalize $(\mathbf{Y}_\mu)_{\mu \in [3]}$ with respect to $\mathbf{Y}_2$, one first uses Algorithm 3.1

ETNA
Kent State University and
Johann Radon Institute (RICAM)

for[1] $(\mathbf{Y}_1, \mathbf{Y}_2{}^{(\{1\})})$ to compute N_1 as a product of inverse non-negative matrices of type (3.2) and to update the cores $\mathbf{Y}_1 \leftarrow \mathbf{Y}_1 N_1$ and $\mathbf{Y}_2{}^{(\{1\})} \leftarrow N_1^{-1}\mathbf{Y}_2{}^{(\{1\})}$. Next, one uses Algorithm 3.1 for[1] $(\mathbf{Y}_3^T, (\mathbf{Y}_2{}^{(\{1,2\})})^T)$ to compute N_2 as a product of inverse non-negative matrices of type (3.2) and to update the cores $\mathbf{Y}_3^T \leftarrow \mathbf{Y}_3^T N_2$ and $(\mathbf{Y}_2{}^{(\{1,2\})})^T \leftarrow N_2^{-1}(\mathbf{Y}_2{}^{(\{1,2\})})^T$. See also Figure 2.4 for an illustration.

The main effort for the quasi-orthogonalization strategy in Algorithm 3.1 lies in solving (3.3), which can be done using, for example, interior point methods for linear programs [20]. Its worst-case complexity is then given in Lemma 3.39.

LEMMA 3.39 (Computational complexity for Algorithm 3.1). *The worst-case computational complexity for Algorithm 3.1 using interior point methods with I inner iteration steps to solve (3.3) and O outer iteration steps in the while loop is of order $\mathcal{O}(OInk^{3.5})$, where $n := \max\{n_1, n_2\}$.*

Proof. The linear program (3.3) can be solved using an interior point method with worst-case complexity of order $\mathcal{O}(N^{2.5}S)$ per iteration, where N is the number of variables and S is the size of the input of the linear program [20]. Here, $N = k - 1$ and $S \in \mathcal{O}(n_1 k)$. Thus, solving (3.3) for β^+ with I iterations each has a worst-case complexity of $\mathcal{O}(In_1k^{3.5})$. All other operations in Algorithm 3.1 have complexities of $\mathcal{O}(nk^2)$ with $n := \max\{n_1, n_2\}$. Hence, the worst-case overall complexity for Algorithm 3.1 is of order $\mathcal{O}(OInk^{3.5})$ using interior point methods. ∎

REMARK 3.40 (Solving micro-steps as quadratic programs). Each micro-step problem (1.3) can be reformulated as a quadratic program of the form $\min_x x^TQx + q^Tx$ s.t. $x \geq 0$ by writing the norm of the residual as inner products. For the ν-th micro-step (1.3), x corresponds to the vectorization of $\mathbf{X}_\nu$, whereas the symmetric operator Q and the vector q result from suitable contractions and reshaping of $\mathcal{A}$, $\mathbf{X}_{\neq\nu}$, and $\mathbf{B}$. Due to this, $Q_{i,:} = (Q_{:,i})^T = \mathbf{0}$ implies that also $q_i = 0$ holds true, and thus not only is the quadratic problem unbounded in $x_i \geq 0$, but also its target function is independent of such entries. In the case that $\mathcal{A}$ is regular, as we consider in Section 4, this also implies that the corresponding tensor $\mathbf{X}$ is independent of such x_i, which can therefore be chosen arbitrarily, but non-negative. For this reason, we restrict x, Q, and q to indices i with $Q_{i,:} \neq \mathbf{0}$. In Section 4, for instance, the old values x_i are kept for all i with $Q_i = \mathbf{0}$.

REMARK 3.41 (On Lemma 3.39). Based on Remark 3.40, solving one micro-step (2.4) for $\mathbf{X}_\mu \in \mathbb{R}^{k\times n\times k}_{\geq 0}$ using an interior point method for quadratic programs with J iterations has a worst-case complexity of $\mathcal{O}(Jn^3k^{10})$ [49]; whereas applying Algorithm 3.1 to $\mathbf{X}_\mu{}^{(\{1,2\})} \in \mathbb{R}^{kn\times k}_{\geq 0}$ with I inner and O outer iterations has a worst-case complexity of $\mathcal{O}(OInk^{4.5})$. Therefore, there seems to be negligible additional cost for Algorithm 3.1 when comparing the worst-case complexities for moderate numbers of iterations. In practice, one observes that including the quasi-orthogonalization strategy in Algorithm 3.1 may further affect the overall runtime of the alternating optimization Algorithm 2.2 (cf. Sections 4.2.6 and 4.3.5). This could result from the fact that the quasi-orthogonalization can increase the search space for each micro-step (2.4). For this reason, the overall additional runtime seems to depend, in particular, on the problem (2.3) itself and the range improvement achieved by the quasi-orthogonalization.

4. Numerical experiments. In this section, we analyze the effect of using the quasi-orthogonalization strategy in Algorithm 3.1 (`quasi-ortho`) within the vanilla alternating non-negative least-squares method in Algorithm 2.2 (`TT-ANLS`) instead of the classical diagonal

[1] Note that here we identify the first core $\mathbf{Y}_1 \in \mathbb{R}^{1\times n\times k}_{\geq 0}$ and the last core $\mathbf{Y}_3 \in \mathbb{R}^{k\times n\times 1}_{\geq 0}$ with their unfoldings $\mathbf{Y}_1{}^{(\{1,2\})} \in \mathbb{R}^{n\times k}_{\geq 0}$ and $\mathbf{Y}_3{}^{(\{1\})} \in \mathbb{R}^{k\times n}_{\geq 0}$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

strategy in Algorithm 2.1 (`diag`) in numerical experiments. We also compare this for the heuristic extrapolation algorithm with restarts in Algorithm A.1 (`TT-HER`). To do so, we focus on two different types of problem. First, we test our method for the non-negative TT approximation of non-negative symmetric polynomials in Section 4.2, for which we provide exact non-negative TT factorizations in Lemma 4.1. In particular, these allow one to modify the dimension, mode size, and factorization size of the target tensors. Second, we analyze our method for the non-negative TT approximation of a particular class of high-dimensional probability distributions in Section 4.3, which have practical applications to modeling tumor progression [37]. Unless otherwise specified, the following settings are used for all experiments.

4.1. Settings. For the tensor-train-related arithmetics, the TT toolbox [33] is used. In Algorithms 2.2 and A.1, each micro-step (1.3) is reformulated as a quadratic program and solved with the MATLAB solver `quadprog` [41][2] as described in Remark 3.40. In Algorithm 3.1, the MATLAB solver `linprog` [41] with at most 10 outer iterations is used to solve (3.3). All experiments are repeated 30 times starting from different randomly generated initial tensors $\mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]}) \in \mathbb{R}_{\geq 0}^{n_1\times\cdots\times n_d}$ with uniformly distributed $\mathbf{X}_\mu = \texttt{rand}(k_\mu, n_\mu, k_{\mu+1})$ for all $\mu \in [d]$. Algorithms 2.1 and 3.1 are abbreviated by `diag` and `quasi-ortho` (in short as `q-o`), and Algorithms 2.2 and A.1 are abbreviated by `TT-ANLS` and `TT-HER`, respectively. In Section 4.3, the `DMRG` solver from the TT toolbox [33] is used to assess bounds on the approximation errors. The `DMRG` solver allows one to solve linear systems of equations approximately within the tensor-train format *without non-negativity constraints* on $\mathbf{X}_\mu$ or additional constraints. To compare relative errors, residuals, and runtimes for all 30 random initial values, geometric means and corresponding variances σ^2 are considered. To visualize the relative errors and residuals of all 30 runs, scatter-like button plots [23] are displayed. These combine overlapping points into larger disks, allowing one to display the corresponding values and their frequencies at the same time. All values are rounded to two significant digits.

4.2. Non-negative approximation of non-negative symmetric polynomials. The first examples on which we focus are non-negative symmetric polynomials $p(x) = (\sum_{\mu=1}^d x_\mu)^r$ of degree r with $x_\mu \in \mathcal{X} \subseteq \mathbb{R}_{\geq 0}$ for all $\mu \in [d]$. In contrast to non-negative tensor factorizations defined by randomly generated non-negative cores, such a polynomial allows one to control the range of its entries, its minimum, and maximum explicitly.

4.2.1. Exact non-negative TT factorization of $\mathbf{P}$. We prove that the corresponding d-dimensional tensor $\mathbf{P}$ has a non-negative TT factorization.

LEMMA 4.1 (Non-negative TT factorization of $\mathbf{P}$). *The non-negative symmetric polynomial tensor $\mathbf{P} \in \mathbb{R}_{\geq 0}^{\mathcal{X}^d}$ of degree r defined by $\mathbf{P}_x := (\sum_{\mu=1}^d x_\mu)^r$ for all $x_\mu \in \mathcal{X} \subseteq \mathbb{R}_{\geq 0}$ has a non-negative TT factorization $(\mathbf{P}_\mu)_{\mu\in[d]}$ of size $(1, r+1, \ldots, r+1, 1)$ with*

$$(\mathbf{P}_1)_{1,x_1,\ell_1} := x_1^{r-\ell_1}\sqrt{\binom{r}{\ell_1}}, \qquad (\mathbf{P}_d)_{\ell_{d-1},x_d,1} := x_d^{\ell_{d-1}-1}\sqrt{\binom{r}{\ell_{d-1}}},$$

$$(\mathbf{P}_\mu)_{\ell_{\mu-1},x_\mu,\ell_\mu} := \begin{cases} 0 & \ell_{\mu-1} < \ell_\mu, \\ x_\mu^{\ell_{\mu-1}-\ell_\mu}\sqrt{\binom{r-\ell_\mu-1}{\ell_{\mu-1}-\ell_\mu}\binom{\ell_{\mu-1}-1}{\ell_{\mu-1}+1}} & \text{otherwise,} \end{cases} \quad \forall\, \mu \in \{2,\ldots,d-1\},$$

for all $x_\nu \in \mathcal{X}$, $\ell_\nu \in [r+1]$, and $\nu \in [d-1]$.

Proof. See Appendix C. ∎

[2] Here, Q and q from Remark 3.40 are rescaled such that $Q_{\max} := \max_{i,j} |Q_{i,j}| = 1$ holds true, if $Q_{\max} < 1$, before calling `quadprog`.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

For this reason, one already knows that $\mathbf{P}$ can be factorized using non-negative tensor trains of sizes bounded by $k = r + 1$, where r denotes its degree.

4.2.2. Problem statement. In the following experiments, we choose symmetric polynomials $\mathbf{P}$ from Section 4.2.1 over $\mathcal{X} = \{0, 1/n - 1, \dots, 1\}$ with $n \in \{2, 3, 5, 10\}$, of degree $r \in \{4, 7\}$, and dimension $d \in \{20, 40, \dots, 120\}$. Then, the goal is to solve (2.3) for $\mathcal{A} := \mathrm{Id}$, $\mathbf{B} := \mathbf{P}$, $k := r + 1$, and $\tau_k(\cdot)$ identifying the TT format, i.e., to find

$$(\mathbf{X}_\mu^+)_{\mu\in[d]} \in \operatorname*{argmin}_{(\mathbf{X}_\mu)_{\mu\in[d]}} \|\mathbf{X} - \mathbf{P}\|_F \quad \text{s.t. } \mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]}),\ \mathbf{X}_\mu \geq 0\ \ \forall\, \mu \in [d]. \tag{4.1}$$

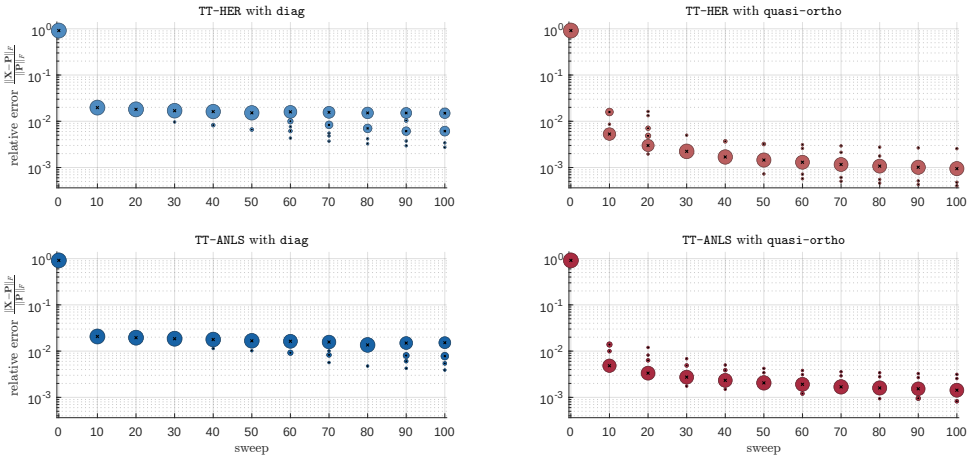
4.2.3. Quality of approximation for $\mathbf{P}$ of degree $r = 4$. We start with the experiments for $\mathbf{P}$ of degree $r = 4$. Table 4.1 shows the geometric means of the relative errors $\|\mathbf{X} - \mathbf{P}\|_F / \|\mathbf{P}\|_F$ for $\mathbf{P}$ as in Section 4.2.2 after 25 sweeps of `TT-ANLS` or `TT-HER` with `diag` or `quasi-ortho` in comparison. The smallest geometric means for each combination are printed in bold. There are no drastic differences between `TT-ANLS` and `TT-HER` comparing the results for `diag` and `qusi-ortho` separately. In turn, comparing the relative errors of `diag` and `quasi-ortho`, it is noticeable that the errors with `quasi-ortho` are lower in all cases considered here. Especially for small mode sizes $n = 2$, the error is reduced by about one order of magnitude. For mode sizes $n \in \{3, 5\}$, such a reduction of the error can still be observed for higher dimensions. For $n = 10$, however, the errors are only reduced by a factor of 2 to 3 on average. These observations apply to both the `TT-ANLS` and `TT-HER` algorithms. The experiments therefore suggest that `TT-ANLS` and `TT-HER` benefit similarly from the use of `quasi-ortho`.

TABLE 4.1
Geometric means of relative errors $\|\mathbf{X} - \mathbf{P}\|_F / \|\mathbf{P}\|_F$ *for* (4.1) *with* $\mathcal{X} = \{0, 1/n - 1, \dots, 1\}$, $n \in \{2, 3, 5, 10\}$, *and* $r = 4$ *after* 25 *sweeps of Algorithm 2.2 or Algorithm A.1 for* 30 *random initial values and geometric variances* $\sigma^2 \in [1.0, 2.0]$.

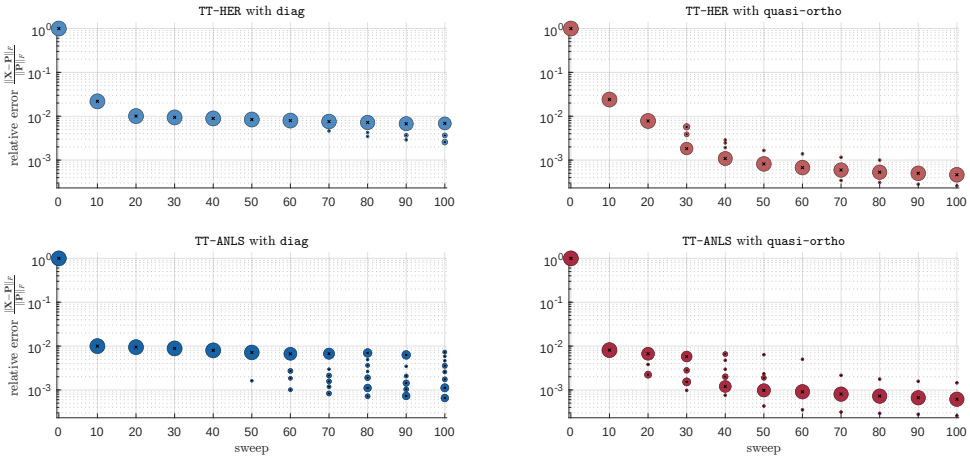
n			$d = 20$	$d = 40$	$d = 60$	$d = 80$	$d = 100$	$d = 120$
2	HER	diag	$1.1\cdot10^{-2}$	$1.6\cdot10^{-2}$	$1.8\cdot10^{-2}$	$1.8\cdot10^{-2}$	$1.7\cdot10^{-2}$	$1.7\cdot10^{-2}$
		q-o	$4.9\cdot10^{-4}$	$\mathbf{2.7\cdot10^{-3}}$	$\mathbf{3.1\cdot10^{-3}}$	$\mathbf{3.4\cdot10^{-3}}$	$\mathbf{2.9\cdot10^{-3}}$	$3.9\cdot10^{-3}$
	ANLS	diag	$1.4\cdot10^{-2}$	$1.8\cdot10^{-2}$	$2.0\cdot10^{-2}$	$2.0\cdot10^{-2}$	$1.9\cdot10^{-2}$	$1.7\cdot10^{-2}$
		q-o	$\mathbf{4.2\cdot10^{-4}}$	$2.8\cdot10^{-3}$	$3.4\cdot10^{-3}$	$3.8\cdot10^{-3}$	$3.2\cdot10^{-3}$	$\mathbf{3.5\cdot10^{-3}}$
3	HER	diag	$6.9\cdot10^{-3}$	$5.7\cdot10^{-3}$	$5.6\cdot10^{-3}$	$9.7\cdot10^{-3}$	$1.4\cdot10^{-2}$	$1.3\cdot10^{-2}$
		q-o	$1.6\cdot10^{-3}$	$\mathbf{2.9\cdot10^{-3}}$	$\mathbf{2.7\cdot10^{-3}}$	$3.0\cdot10^{-3}$	$3.9\cdot10^{-3}$	$4.7\cdot10^{-3}$
	ANLS	diag	$7.1\cdot10^{-3}$	$5.8\cdot10^{-3}$	$6.8\cdot10^{-3}$	$7.2\cdot10^{-3}$	$8.0\cdot10^{-3}$	$9.1\cdot10^{-3}$
		q-o	$\mathbf{1.1\cdot10^{-3}}$	$3.0\cdot10^{-3}$	$2.8\cdot10^{-3}$	$\mathbf{2.5\cdot10^{-3}}$	$\mathbf{2.6\cdot10^{-3}}$	$\mathbf{2.8\cdot10^{-3}}$
5	HER	diag	$3.6\cdot10^{-3}$	$3.3\cdot10^{-3}$	$1.4\cdot10^{-2}$	$1.3\cdot10^{-2}$	$1.1\cdot10^{-2}$	$1.0\cdot10^{-2}$
		q-o	$\mathbf{2.0\cdot10^{-3}}$	$\mathbf{2.7\cdot10^{-3}}$	$2.1\cdot10^{-3}$	$3.1\cdot10^{-3}$	$4.6\cdot10^{-3}$	$5.2\cdot10^{-3}$
	ANLS	diag	$4.0\cdot10^{-3}$	$3.4\cdot10^{-3}$	$3.5\cdot10^{-3}$	$7.5\cdot10^{-3}$	$8.6\cdot10^{-3}$	$8.2\cdot10^{-3}$
		q-o	$\mathbf{2.0\cdot10^{-3}}$	$3.0\cdot10^{-3}$	$\mathbf{1.9\cdot10^{-3}}$	$\mathbf{2.8\cdot10^{-3}}$	$\mathbf{2.7\cdot10^{-3}}$	$\mathbf{3.5\cdot10^{-3}}$
10	HER	diag	$3.3\cdot10^{-3}$	$1.2\cdot10^{-2}$	$1.3\cdot10^{-2}$	$1.1\cdot10^{-2}$	$9.6\cdot10^{-3}$	$8.3\cdot10^{-3}$
		q-o	$\mathbf{1.9\cdot10^{-3}}$	$\mathbf{1.6\cdot10^{-3}}$	$2.4\cdot10^{-3}$	$3.1\cdot10^{-3}$	$4.5\cdot10^{-3}$	$5.5\cdot10^{-3}$
	ANLS	diag	$3.1\cdot10^{-3}$	$3.0\cdot10^{-3}$	$9.9\cdot10^{-3}$	$1.0\cdot10^{-2}$	$6.3\cdot10^{-3}$	$7.9\cdot10^{-3}$
		q-o	$2.0\cdot10^{-3}$	$1.7\cdot10^{-3}$	$\mathbf{1.7\cdot10^{-3}}$	$\mathbf{2.8\cdot10^{-3}}$	$\mathbf{3.4\cdot10^{-3}}$	$\mathbf{3.3\cdot10^{-3}}$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

4.2.4. Convergence behavior. Figure 4.1 shows semi-logarithmic button plots (see Section 4.1 or [23]) for the relative error as a function of the sweeps for `diag` on the left (in blue), `quasi-ortho` on the right (in bordeaux), `TT-HER` above, and `TT-ANLS` below for dimension $d = 100$. Figure 4.1(a) shows the decay for mode size $n = 2$ and Figure 4.1(b) for $n = 10$. In both Figures 4.1(a) and 4.1(b) using `diag`, the error always decreases rapidly to approximately 10^{-2} in the first 10–20 sweeps. After 50–80 sweeps, there are some samples for which the errors start to further decrease, but slowly. Using `quasi-ortho` instead, the relative errors also drop after 10 sweeps but continue to decrease, though more slowly, in all cases. The relative error for `TT-HER` with `quasi-ortho` decreases slightly slower but more evenly than for `TT-ALS`. Concluding, Figure 4.1 suggests that using `quasi-ortho` instead of `diag` helps to prevent stagnation and to speed up convergence for small and larger mode sizes. However, as is typical for such alternating minimization strategies, a large number of sweeps may be required to achieve a higher approximation accuracy.



(a) $n = 2$ and $d = 100$



(b) $n = 10$ and $d = 100$

FIG. 4.1. *Semi-logarithmic plots of the relative error* $\|\mathbf{X} - \mathbf{P}\|_F / \|\mathbf{P}\|_F$ *as a function of the sweeps for* (4.1) *with* $\mathcal{X} = \{0, 1/n - 1, \ldots, 1\}$, $d = 100$, $n \in \{2, 10\}$, *and* $r = 4$ *using Algorithm 2.2 or Algorithm A.1 for* 30 *random initial values.*

ETNA
Kent State University and
Johann Radon Institute (RICAM)

4.2.5. Quality of approximation for $\mathbf{P}$ of degree $r = 7$. In the same setting as in Section 4.2.3, we repeat the experiments for the extreme mode sizes $n = 2$ and $n = 10$, but with $\mathbf{P}$ of degree $r = 7$, i.e., for $k = 8$ in (4.1). Table 4.2 shows the geometric means of the relative errors $\|\mathbf{X} - \mathbf{P}\|_F / \|\mathbf{P}\|_F$. Again, the smallest mean is written in bold for each combination. The results in Table 4.2 behave quite similarly to those in Table 4.1 for $r = 4$, i.e., for $k = 5$. Comparing `TT-HER` and `TT-ANLS`, oftentimes `TT-HER` allows for slightly further reduction of the relative error on average. Comparing `diag` and `quasi-ortho` instead, the relative errors for $n = 2$ are approximately one order of magnitude smaller using `quasi-ortho` compared to `diag`. For $n = 10$ and $d \leq 100$, the relative errors are on average again about half as large when using `quasi-ortho` instead of `diag`. When increasing the dimension d, this effect of `quasi-ortho` seems to be further pronounced. Overall, both `TT-ANLS` and `TT-HER` seem to benefit from the use of `quasi-ortho` again. The combination of `quasi-ortho` and `TT-HER` nearly always leads to the smallest relative error on average here. However, this is also associated with a higher runtime, as can be observed in the following section.

TABLE 4.2
Geometric means of relative errors $\|\mathbf{X} - \mathbf{P}\|_F / \|\mathbf{P}\|_F$ *for* (4.1) *with* $\mathcal{X} = \{0, 1/n - 1, \ldots, 1\}$, $n \in \{2, 10\}$, *and* $r = 7$ *after* 25 *sweeps of Algorithm* 2.2 *or Algorithm* A.1 *for* 30 *random initial values and geometric variances* $\sigma^2 \in [1.0, 2.3]$.

n			$d = 20$	$d = 40$	$d = 60$	$d = 80$	$d = 100$	$d = 120$
2	HER	diag	$1.1 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$1.0 \cdot 10^{-2}$	$9.6 \cdot 10^{-3}$	$9.6 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$
		q-o	$4.5 \cdot 10^{-4}$	$\mathbf{4.2 \cdot 10^{-3}}$	$\mathbf{5.5 \cdot 10^{-3}}$	$\mathbf{6.1 \cdot 10^{-3}}$	$\mathbf{5.7 \cdot 10^{-3}}$	$\mathbf{5.6 \cdot 10^{-3}}$
	ANLS	diag	$1.4 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$
		q-o	$\mathbf{2.8 \cdot 10^{-4}}$	$5.6 \cdot 10^{-3}$	$8.0 \cdot 10^{-3}$	$7.4 \cdot 10^{-3}$	$7.0 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$
10	HER	diag	$2.1 \cdot 10^{-3}$	$2.7 \cdot 10^{-3}$	$5.8 \cdot 10^{-3}$	$1.0 \cdot 10^{-2}$	$1.2 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$
		q-o	$\mathbf{1.5 \cdot 10^{-3}}$	$\mathbf{1.5 \cdot 10^{-3}}$	$\mathbf{1.5 \cdot 10^{-3}}$	$\mathbf{1.6 \cdot 10^{-3}}$	$2.3 \cdot 10^{-3}$	$\mathbf{2.8 \cdot 10^{-3}}$
	ANLS	diag	$2.1 \cdot 10^{-3}$	$3.1 \cdot 10^{-3}$	$2.8 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	$3.7 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$
		q-o	$\mathbf{1.5 \cdot 10^{-3}}$	$2.6 \cdot 10^{-3}$	$2.2 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$	$\mathbf{2.2 \cdot 10^{-3}}$	$6.6 \cdot 10^{-3}$

4.2.6. Runtimes. In addition to the worst-case complexities of Algorithm 3.1 in Lemma 3.39, Table 4.3 presents the geometric means of practical runtimes for `diag` in seconds and the quotient of the means for all other strategies. For the simple approximation problem (4.1), the runtimes of `diag-HER` are 10% to 30% larger than those of `diag-ANLS`. The quotient is slightly increasing in the dimension d and decreasing in n and r. Further, the runtimes of `quasi-ortho-ANLS` and `quasi-ortho-HER` are 1.6–8.2 and 1.8–8.8 times higher that those of `diag-ANLS`, respectively. But note that already after 10 sweeps the relative error is oftentimes smaller than the final one using `diag`, in particular, for $n = 2$. Both quotients seem to decrease in n and r, but vary in the dimension d. As expected, also `quasi-ortho-HER` has longer runtimes than `quasi-ortho-ANLS` but the main extra time is needed for `quasi-ortho` compared to `diag`. Thus, these observations agree with the theoretical considerations from Remark 3.41.

4.2.7. Relaxation of the non-negativity constraint in Algorithm 3.1. As mentioned in Remark 3.37, the non-negativity constraint in (3.3) could be relaxed by replacing the constraint $Y\beta \geq 0$ with $Y\beta \geq -\varepsilon$, see (3.4), and then projecting Y and Z back to the non-negative orthant. Here the hope is to further increase the non-negative range of Y by relaxing the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

TABLE 4.3
Geometric means and quotients of runtimes in seconds for (4.1) with $\mathcal{X} = \{0, 1/n - 1, \dots, 1\}$, $n \in \{2, 3, 5, 10\}$, and $r \in \{4, 7\}$ using Algorithm 2.2 or Algorithm A.1 with 25 sweeps and 30 random initial values and geometric variances $\sigma^2 \in [1.0, 1.1]$.

n	r		$d = 20$	$d = 40$	$d = 60$	$d = 80$	$d = 100$	$d = 120$
2	4	`diag-A` (s)	7.6	18	28	38	51	66
		`q-o-A/diag-A`	8.1	7.6	6.6	6.0	6.0	5.7
		`diag-H/diag-A`	1.1	1.2	1.3	1.3	1.3	1.3
		`q-o-H/diag-A`	8.2	7.9	7.2	7.2	8.1	8.8
	7	`diag-A` (s)	16	35	57	80	97	139
		`q-o-A/diag-A`	7.1	6.2	5.2	5.4	5.4	4.7
		`diag-H/diag-A`	1.1	1.1	1.1	1.1	1.2	1.2
		`q-o-H/diag-A`	7.4	6.5	5.8	6.3	7.4	6.4
10	4	`diag-A` (s)	32	67	104	147	198	252
		`q-o-A/diag-A`	2.4	2.3	2.6	2.8	2.4	2.4
		`diag-H/diag-A`	1.1	1.1	1.1	1.1	1.1	1.2
		`q-o-H/diag-A`	2.4	3.3	3.7	4.1	3.5	3.5
	7	`diag-A` (s)	101	215	342	484	614	754
		`q-o-A/diag-A`	1.8	1.6	1.6	1.6	1.6	1.6
		`diag-H/diag-A`	1.1	1.1	1.1	1.1	1.1	1.1
		`q-o-H/diag-A`	1.8	2.1	2.1	2.1	2.1	2.0

constraint as above. Table D.1 in Appendix D shows the geometric means of the relative errors for $\varepsilon \in \{0, 10^{-16}, 10^{-12}, 10^{-8}, 10^{-4}, 1\}$ and all combinations of $n \in \{2, 10\}$ and $d \in \{20, 40, \dots, 120\}$ using `TT-HER` and `TT-ANLS` with `quasi-ortho`, respectively. For $\varepsilon \leq 10^{-8}$, the respective relative errors are in the same order of magnitude. For $\varepsilon > 10^{-8}$ the relative error increases significantly. In conclusion, relaxing the non-negativity constraint in (3.3) using (3.4) might help to further speed up convergence if $\varepsilon > 0$ is chosen properly. The geometric variances σ^2 over all 30 random initializations are typically larger for $\varepsilon > 0$. Note that a proper choice for $\varepsilon \geq 0$ may depend on the problem, the micro-step solver, the solver used for (3.4), and also the initial factorization chosen.

4.3. Non-negative approximation for probability distributions. One particular application we have in mind is the non-negative approximation of high-dimensional probability distributions that are defined as a solution of a linear system. Here, we focus on one that arises in the context of tumor progression modeling via high-dimensional Markov chains [37].

4.3.1. The model. The searched distribution $\mathbf{X}^\star$ over the discrete state space $\mathcal{S} := \{0, 1\}^d$ is defined as the solution of the linear system

$$\mathcal{A}(\mathbf{X}) = \mathbf{B},$$

where the operator $\mathcal{A} \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$ and the right-hand side

$$\mathbf{B} := \bigotimes_{\nu=1}^{d} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^{\mathcal{S}}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

both have low-rank representations with

$$\mathcal{A} := \bigotimes_{\nu=1}^{d} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \sum_{\mu=1}^{d} \bigotimes_{\nu=1}^{\mu-1} \begin{bmatrix} 1 & 0 \\ 0 & \Theta_{\mu,\nu} \end{bmatrix} \otimes \begin{bmatrix} \Theta_{\mu,\mu} & 0 \\ -\Theta_{\mu,\mu} & 0 \end{bmatrix} \otimes \bigotimes_{\nu=\mu+1}^{d} \begin{bmatrix} 1 & 0 \\ 0 & \Theta_{\mu,\nu} \end{bmatrix}$$

for some parameters $\Theta \in \mathbb{R}^{d \times d}_{>0}$. Note that, due to the model construction, the mode sizes of $\mathbf{X}^\star$ are all equal to $n = 2$. This suggests an improvement of `quasi-ortho` over `diag` in particular for higher dimensions, as seen in Section 4.2. Further, the dimension d corresponds to the number of genetic events, e.g., mutations, considered in the Markov chain. There are up to 800 genes known to be involved in tumor progression [2], the mutations of which could be included as events in a model. Typically, not all of these are relevant at the same time, but higher dimensions $d \gtrsim 100$ are needed for comprehensive models. In the following, we use synthetic parameters Θ generated according to [8, Section 4.1, (B1)] and choose the dimensions d to be a multiple of 8 to be consistent with [8].

4.3.2. Problem statement. As $\mathbf{X}^\star$ is a probability distribution, $\mathbf{X}^\star \in [0,1]^{\mathcal{S}}$ and $\langle \mathbf{1}_{\mathcal{S}}, \mathbf{X}^\star \rangle_F = 1$ hold true. In [8] it was already demonstrated that $\mathbf{X}^\star$ can be well approximated using low-rank tensors (at least, for a specific class of parameters Θ). In order to allow for further interpretation of its low-rank approximation, we also want its approximation to fulfill both conditions. This results in the task to find

$$(\mathbf{X}_\mu^+)_{\mu\in[d]} \in \operatorname*{argmin}_{(\mathbf{X}_\mu)_{\mu\in[d]}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F \quad \text{s.t. } \mathbf{X} = \tau_k((\mathbf{X}_\mu)_{\mu\in[d]}), \quad \mathbf{X}_\mu \geq 0 \ \ \forall \mu \in [k], \ \langle \mathbf{C}, \mathbf{X} \rangle_F = 1, \tag{4.2}$$

where

$$\mathbf{C} := \mathbf{1}_{\mathcal{S}} = \bigotimes_{\nu=1}^{d} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in \mathbb{R}^{\mathcal{S}}$$

and k denotes the target size. Note that the additional constraint $\langle \mathbf{C}, \mathbf{X} \rangle_F = 1$ can be reformulated as an equality constraint for $\mathbf{X}_\nu$ with fixed $\mathbf{X}_{\neq\nu}$ and therefore can be included in each micro-step (2.4). The resulting micro-step can then also be solved using a quadratic program solver like `quadprog` [41].

4.3.3. Quality of approximation. Similar to Sections 4.2.3 and 4.2.5, the geometric means of the relative residuals $(\|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F)/\|\mathbf{B}\|_F$ for dimensions $d \in \{24, 32, \dots, 72\}$ and $k \in \{5, 8\}$ are summarized in Table 4.4. The last column in each part of the table gives the geometric mean of the relative residuals achieved by `DMRG` [33] to stagnation without non-negativity constraints on the cores $(\mathbf{X}_\mu)_{\mu\in[d]}$. For this reason, the residuals for `DMRG` can be regarded as an approximate lower bound for the expected residuals of the other methods. However, as `DMRG` in general ensures neither non-negativity of the cores $(\mathbf{X}_\mu)_{\mu\in[d]}$ nor the equality condition $\langle \mathbf{C}, \mathbf{X} \rangle_F = 1$, `DMRG` generally does not allow one to solve (2.3) or (4.2).

For $k = 5$, one observes that the relative residuals achieved using `TT-ANLS` and `TT-HER` are typically in the same order of magnitude as those achieved by `DMRG`, i.e., one cannot expect to get closer to the solution without increasing k. By increasing k to $k = 8$, `DMRG`, which ignores both the non-negativity and equality constraint, achieves a relative residual that is one order smaller for $d \geq 48$. However, it is important to note that this does not necessarily imply that the solution $\mathbf{X}^\star$ can be approximated non-negatively with a residual of the same order. For both ranks $k \in \{5, 8\}$, there is a trend that the solution $\mathbf{X}^\star$ can be better approximated by low-rank or non-negative factorizations for higher dimensions d. This observation is consistent

ETNA
Kent State University and
Johann Radon Institute (RICAM)

with [8] and seems to be due to the problem statement in Section 4.3.2. A comparison of TT-HER and TT-ANLS shows only slight differences and no clear trend as to which method is superior. Comparing diag and quasi-ortho, quasi-ortho leads to a lower geometric mean of the relative residuals in all cases. In most cases the differences are small, but in some cases are up to an order of magnitude; see, e.g., $k = 8$ and $d = 40$ using TT-ALS. The lowest relative residuals on average are always achieved by TT-ANLS with quasi-ortho except for $d = 64$ and $k = 8$.

TABLE 4.4
Geometric means of relative residuals $(\|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F)/\|\mathbf{B}\|_F$ *for* (4.2) *with* $n = 2$ *and* $k \in \{5, 8\}$*, using Algorithms 2.2 and A.1, and DMRG [33] with* 30 *random initial values and geometric variances* $\sigma^2 \in [1.0, 1.3]$.

d	$k = 5$				
	TT-HER		TT-ANLS		(DMRG)
	diag	q-o	diag	q-o	
24	$4.0 \cdot 10^{-3}$	$2.8 \cdot 10^{-3}$	$3.5 \cdot 10^{-3}$	$\mathbf{2.6 \cdot 10^{-3}}$	$(2.1 \cdot 10^{-3})$
32	$2.2 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$	$\mathbf{1.7 \cdot 10^{-3}}$	$(1.6 \cdot 10^{-3})$
40	$1.7 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$	$\mathbf{1.2 \cdot 10^{-3}}$	$(7.5 \cdot 10^{-4})$
48	$1.2 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$1.3 \cdot 10^{-3}$	$\mathbf{9.0 \cdot 10^{-4}}$	$(7.8 \cdot 10^{-4})$
56	$1.2 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$1.3 \cdot 10^{-3}$	$\mathbf{9.0 \cdot 10^{-4}}$	$(7.8 \cdot 10^{-4})$
64	$8.3 \cdot 10^{-4}$	$7.5 \cdot 10^{-4}$	$8.8 \cdot 10^{-4}$	$\mathbf{7.4 \cdot 10^{-4}}$	$(3.1 \cdot 10^{-4})$
72	$9.0 \cdot 10^{-4}$	$7.7 \cdot 10^{-4}$	$6.6 \cdot 10^{-4}$	$\mathbf{6.5 \cdot 10^{-4}}$	$(2.3 \cdot 10^{-4})$

d	$k = 8$				
	TT-HER		TT-ANLS		(DMRG)
	diag	q-o	diag	q-o	
24	$2.2 \cdot 10^{-3}$	$1.4 \cdot 10^{-3}$	$3.2 \cdot 10^{-3}$	$\mathbf{1.4 \cdot 10^{-3}}$	$(1.1 \cdot 10^{-3})$
32	$1.0 \cdot 10^{-3}$	$6.3 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$	$\mathbf{5.7 \cdot 10^{-4}}$	$(5.2 \cdot 10^{-4})$
40	$9.8 \cdot 10^{-4}$	$4.7 \cdot 10^{-4}$	$7.3 \cdot 10^{-3}$	$\mathbf{4.0 \cdot 10^{-4}}$	$(1.0 \cdot 10^{-4})$
48	$4.2 \cdot 10^{-4}$	$3.2 \cdot 10^{-4}$	$6.7 \cdot 10^{-4}$	$\mathbf{2.5 \cdot 10^{-4}}$	$(4.8 \cdot 10^{-5})$
56	$5.1 \cdot 10^{-4}$	$3.2 \cdot 10^{-4}$	$4.3 \cdot 10^{-4}$	$\mathbf{2.7 \cdot 10^{-4}}$	$(3.2 \cdot 10^{-5})$
64	$3.1 \cdot 10^{-4}$	$\mathbf{2.7 \cdot 10^{-4}}$	$6.2 \cdot 10^{-4}$	$3.4 \cdot 10^{-4}$	$(2.8 \cdot 10^{-5})$
72	$4.8 \cdot 10^{-4}$	$2.7 \cdot 10^{-4}$	$2.6 \cdot 10^{-4}$	$\mathbf{2.2 \cdot 10^{-4}}$	$(1.7 \cdot 10^{-5})$

4.3.4. Convergence behavior. In Figure 4.2, the semi-logarithmic button plots again display the relative residuals as a function of the sweeps for $d = 48$ with $k = 5$. The results using diag are on the left (in blue), quasi-ortho on the right (in bordeaux), TT-HER above, and TT-ANLS below. In all four cases, the relative residual drops fast during the first 10 sweeps and then reduces only quite slowly. Using TT-HER compared to TT-ALS, the fast reduction during the first sweeps is slower for most initial values but reaches similar values after about 20 sweeps. Using diag in TT-HER or TT-ALS, there are some initial values for which the relative residual cannot be reduced to $1 \cdot 10^{-3}$ but stagnates after about 10 sweeps between $3 \cdot 10^{-3}$ and $8 \cdot 10^{-3}$. Using TT-ALS with quasi-ortho, the relative residuals behave similarly for all initial values. The convergence behaviors are similar for other dimensions d and therefore are not shown here.

4.3.5. Runtimes. Similar to Table 4.3, Table 4.5 also lists the geometric means of the runtimes in seconds for all experiments. In contrast to Section 4.2.6, one observes that using quasi-ortho for (4.2) has a much smaller impact on the runtimes. Moreover,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

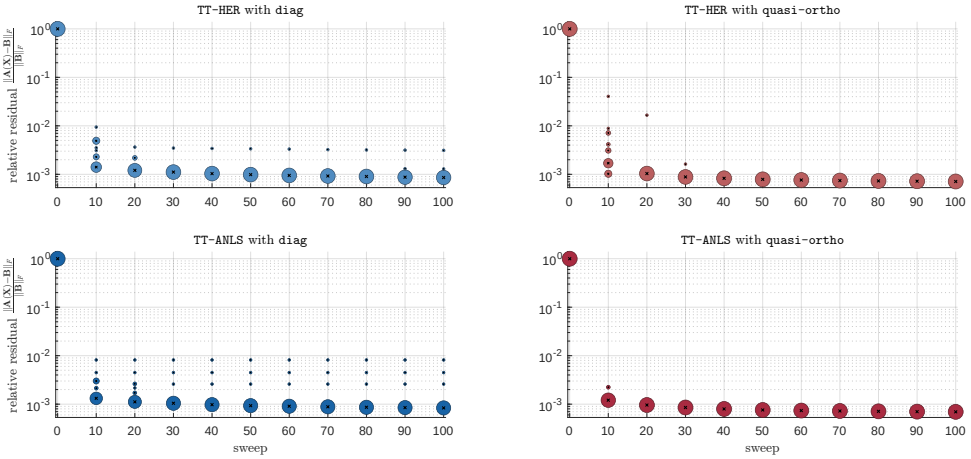


FIG. 4.2. *Semi-logarithmic plots of the relative residual* $(\|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F)/\|\mathbf{B}\|_F$ *as a function of the sweeps for* (4.2) *with* $n = 2$, $d = 48$, *and* $k = 5$ *using Algorithms* 2.2 *and* A.1 *with* 30 *random initial values.*

this impact appears to decrease with increasing dimension d and size k. For `TT-ANLS`, using `quasi-ortho` instead of `diag` has almost no effect on the computational time for $d \geq 56$. Using `TT-HER` instead of `TT-ANLS` approximately doubles the runtimes. The use of `quasi-ortho` can further reduce the runtime compared to `diag` for higher dimensions; see, e.g., $d = 72$ and $k \in \{5.8\}$ for both `TT-ANLS` and `TT-HER`. This supports the conjecture that the cost increase due to `quasi-ortho` is less significant for more complex problems. When comparing the times required for 25 sweeps with those for 100 sweeps (not shown here), the latter take approximately four times as long as the former, whereas the ratios between `TT-ANLS` and `TT-HER` as well as `diag` and `quasi-ortho` remain similar.

TABLE 4.5
Geometric means and quotients of runtimes in seconds for (4.2) *with* $n = 2$ *and* $k \in \{5, 8\}$ *using Algorithms* 2.2 *and* A.1 *with* 25 *sweeps for* 30 *random initial values and geometric variances* $\sigma^2 \in [1.0, 1.1]$.

d	$k=5$				$k=8$			
	`TT-ANLS`		`TT-HER`		`TT-ANLS`		`TT-HER`	
	`diag-A` (s)	`q-o-A` / `diag-A`	`diag-H` / `diag-A`	`q-o-H` / `diag-A`	`diag-A` (s)	`q-o-A` / `diag-A`	`diag-H` / `diag-A`	`q-o-H` / `diag-A`
24	114	1.5	1.9	2.5	266	1.4	1.8	2.5
32	318	1.3	1.9	2.3	679	1.2	1.9	2.1
40	654	1.1	1.9	2.0	1431	1.1	1.9	2.1
48	1193	1.1	2.0	2.0	2918	1.0	1.9	1.9
56	2204	1.0	2.0	1.9	5174	1.0	1.9	1.9
64	3308	1.0	2.0	2.0	8382	0.96	2.0	1.9
72	5978	0.99	2.0	1.9	14 608	0.98	2.0	1.9

5. Conclusion. We developed the quasi-orthogonalization method as an intermediate step between two micro-steps in alternating non-negative tensor factorization with the following goal: to improve the expressivity in each non-negative factor by increasing the non-negative ranges of all fixed factors, but necessarily without changing the tensor represented or losing the non-negativity of all factors. To do this, we derived how to obtain such a factorization within the equivalence class representing the same tensor by modifying the current one

ETNA
Kent State University and
Johann Radon Institute (RICAM)

via certain transfer matrices. We focused on a particular, numerically manageable subset of M-matrices, whose elements are inverse non-negative under mild conditions and thus allow for non-negative updates of all factors. Further, we proved that any such transfer matrix, which is not a permuted diagonal matrix, leads to a strict increase in the non-negative range. We proposed a simple strategy for finding such a transfer matrix based on well-known linear programming approaches. Numerical experiments suggest that including this into an alternating non-negative approach indeed reduces the persistence in local minima and improves the convergence properties of the alternating optimization, where the differences, in particular for higher dimensions and small mode sizes, can be significant.

Future work may include answering open questions such as how to find an optimal transfer matrix that allows for a maximal non-negative range, and how to use more general transfer matrices. Likewise, further modifications of alternating optimization methods for tensors may be extended to the non-negative case.

Appendix A. Heuristic extrapolation with restarts (HER) for tensor trains. The HER method from [29] adapted to tensor trains and extended by a quasi-orthogonalization step is given in Algorithm A.1. To simplify the notation, we do not distinguish between the two half-sweeps from $\nu = 1$ to $d-1$ and from $\nu = d$ to 2 in Algorithm A.1. However, this can be done analogously to Algorithm 2.2.

Algorithm A.1: `TT-HER` for (2.3)

Input: $\mathcal{A}$ TT operator, $\mathbf{B}$ TT tensor, $\mathbf{X}^{(0)} = \tau_k((\mathbf{X}_\mu^{(0)})_{\mu\in[d]})$ initial TT tensor with $\mathbf{X}_\mu^{(0)} \geq 0$, $\omega \in (0,1)$, and $1 \leq \gamma \leq \overline{\gamma} \leq \eta$ HER parameter

Output: $\mathbf{X}$ approximate solution of (2.3)

Set $\overline{\omega} = 1$ // Initialize $\overline{\omega}$ (upper bound on ω)
Set $i = 0$
Set $\widehat{\mathbf{X}}_\mu^{(i)} = \mathbf{X}_\mu^{(i)}$ for all $\mu \in [d]$ // Initialize $\widehat{\mathbf{X}}^{(i)} = \tau_k((\widehat{\mathbf{X}}_\mu^{(i)})_{\mu\in[d]})$
while *stop criteria are not fulfilled* **do**
    Set $\widehat{\mathbf{X}}_\mu^{(i+1)} = \widehat{\mathbf{X}}_\mu^{(i)}$ for all $\mu \in [d]$ // Initialize $\widehat{\mathbf{X}}^{(i+1)}$
    for $\nu \in [d]$ **do** // Perform a half-sweep
        Quasi-orthogonalize or normalize $\widehat{\mathbf{X}}^{(i+1)}$ w.r.t. $\widehat{\mathbf{X}}_\nu^{(i)}$
        Solve $\mathbf{X}_\nu^{(i+1)} \in \operatorname{argmin}_{\mathbf{X}_\nu \geq 0} \|\mathcal{A}(\tau_k(\widehat{\mathbf{X}}_{\neq\nu}^{(i+1)}, \mathbf{X}_\nu)) - \mathbf{B}\|_F^2$
        Update $\widehat{\mathbf{X}}_\nu^{(i+1)} \leftarrow \max\{0, (1+\omega)\mathbf{X}_\nu^{(i+1)} - \omega\mathbf{X}_\nu^{(i)}\}$ // Extrapolation
    end
    if $\|\mathcal{A}(\widehat{\mathbf{X}}^{(i+1)}) - \mathbf{B}\|_F^2 > \|\mathcal{A}(\widehat{\mathbf{X}}^{(i)}) - \mathbf{B}\|_F^2$ **then** // Restart
        $\mathbf{X}_\mu^{(i+1)} \leftarrow \mathbf{X}_\mu^{(i)}$ and $\widehat{\mathbf{X}}_\mu^{(i+1)} \leftarrow \mathbf{X}_\mu^{(i+1)}$ for all $\mu \in [d]$
        $\overline{\omega} \leftarrow \omega$ and $\omega \leftarrow \omega/\eta$ // Decrease $\overline{\omega}$ and ω
    else
        $\mathbf{X}_\mu^{(i+1)} \leftarrow \widehat{\mathbf{X}}_\mu^{(i+1)}$ for all $\mu \in [d]$
        $\overline{\omega} \leftarrow \min\{1, \overline{\omega} \cdot \overline{\gamma}\}$ and $\omega \leftarrow \min\{\overline{\omega}, \omega \cdot \gamma\}$ // Increase $\overline{\omega}$ and ω
    end
    Increase $i \leftarrow i + 1$
end
return $\mathbf{X}$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Appendix B. Further lemmas. In the proof of Theorem 3.31, we use the following lemmas.

LEMMA B.1. *For all $p \geq 1$ and $a, b \geq 0$ it holds true that $(a+b)^p \geq a^p + b^p$.*

Proof. If $a = b = 0$ or $p = 1$, the statement is clear. Otherwise, we define $t := a/(a+b) \in [0,1]$ with $(1-t) = b/(a+b) \in [0,1]$. Then $t^p + (1-t)^p \leq t + (1-t) = 1$ holds true and multiplying both sides by $(a+b)^p > 0$ leads to the above result. ∎

LEMMA B.2. *Let $p \geq 1$, $\alpha \in \mathbb{R}^k_{\geq 0}$, and $v^{(1)}, \dots, v^{(k)} \in \mathbb{R}^{n_1}_{\geq 0}$ with $\|v^{(\ell)}\|_p = 1$ for all $\ell \in [k]$. Then $\|\sum_{\ell=1}^k \alpha_\ell v^{(\ell)}\|_p \geq \|\alpha\|_p$ holds true. For $p = 1$, equality holds true.*

Proof. For $k = 1$ or $p = 1$, the statement is trivial. For $k > 1$, the statement follows by induction using the result for $k = 2$:

$$\begin{aligned}\|\alpha_1 v^{(1)} + \alpha_2 v^{(2)}\|_p^p &= \sum_{j=1}^{n_1} (\alpha_1 v_j^{(1)} + \alpha_2 v_j^{(2)})^p \overset{\text{Lem. B.1}}{\geq} \sum_{j=1}^{n_1} (\alpha_1 v_j^{(1)})^p + \sum_{j=1}^{n_1} (\alpha_2 v_j^{(2)})^p \\ &= \alpha_1^p + \alpha_2^p = \|\alpha\|_p^p. \qquad \blacksquare\end{aligned}$$

Appendix C. Proofs. The complete proofs of Theorem 3.5 and Lemma 4.1 are given in this appendix.

Proof of Theorem 3.5. We prove the statements (i) to (iv) of the theorem by constructing $\mathcal{I} \subseteq [k]$ as follows.

(iii) Let Y contain no column which is a multiple of another one. Removing an index $\ell \in [k]$ of an extreme column $Y_{:,\ell}$ would result in $\operatorname{range}_{\geq 0}(Y_{:,\neq \ell}) \subsetneq \operatorname{range}_{\geq 0}(Y)$. Thus, $\mathcal{I}$ must contain all indices of extreme columns. As $Y \neq \mathbf{0}$, there exists at least one extreme column, i.e., $\mathcal{I} \neq \emptyset$. Let $\mathcal{J} := [k] \setminus \mathcal{I}$ denote the set of all non-extreme columns. Then for each $j \in \mathcal{J}$ there exists an $\alpha^{(j)} \geq 0$ such that $Y_{:,j} = \sum_{\ell \neq j} \alpha_\ell^{(j)} Y_{:,\ell}$ and $Y_{:,j} \neq \lambda Y_{:,\ell}$ for all $\lambda \geq 0$ and $\ell \neq j$. Without loss of generality, let $\mathcal{J} = [n]$. We show that $Y_{:,\leq j} \in \operatorname{range}_{\geq 0}(Y_{:,>j})$ by induction over $j \in \mathcal{J}$. For $j = 1$, the statement directly holds true as $Y_{:,j_1}$ is non-extreme. Let $Y_{:,<j} \subseteq \operatorname{range}_{\geq 0}(Y_{:,\geq j})$ hold true with $Y_{:,i} = \sum_{\ell \geq j} \beta_\ell^{(i)} Y_{:,\ell}$ with $\beta^{(i)} \geq 0$ for all $i < j$. Then we have

$$\begin{aligned} Y_{:,j} &= \sum_{i<j} \alpha_i^{(j)} Y_{:,i} + \sum_{\ell>j} \alpha_\ell^{(j)} Y_{:,\ell} \\ &= \left(\sum_{i<j} \alpha_i^{(j)} \beta_j^{(i)}\right) Y_{:,j} + \sum_{\ell>j} \left(\alpha_\ell^{(j)} + \sum_{i<j} \alpha_i^{(j)} \beta_\ell^{(i)}\right) Y_{:,\ell} \\ \iff \quad \left(1 - \sum_{i<j} \alpha_i^{(j)} \beta_j^{(i)}\right) Y_{:,j} &= \sum_{\ell>j} \left(\alpha_\ell^{(j)} + \sum_{i<j} \alpha_i^{(j)} \beta_\ell^{(i)}\right) Y_{:,\ell}. \end{aligned}$$

As $Y \geq 0$ and $\alpha^{(j)}, \beta^{(i)} \geq 0$, directly $1 - \sum_{i<j} \alpha_i^{(j)} \beta_j^{(i)} \geq 0$ follows. If $\sum_{i<j} \alpha_i^{(j)} \beta_j^{(i)} = 1$, then

$$0 = \sum_{\ell>j} \left(\alpha_\ell^{(j)} + \sum_{i<j} \alpha_i^{(j)} \beta_\ell^{(i)}\right) Y_{:,\ell},$$

which implies that $\alpha_\ell^{(j)} = 0$ and $\alpha_i^{(j)} \beta_\ell^{(i)} = 0$ for all $i < j < \ell$. Assume there exists an $i < j$ with $\alpha_i^{(j)} > 0$, then $\beta_\ell^{(i)} = 0$ for all $\ell > j$ and $Y_{:,i} = \beta_j^{(i)} Y_{:,j}$, which

ETNA
Kent State University and
Johann Radon Institute (RICAM)

contradicts the assumption. Thus, $\alpha^{(j)} = 0$ would imply that $Y_{:,j} = \mathbf{0}$, which also contradicts the assumption. For this reason, $1 - \sum_{i<j} \alpha_i^{(j)} \beta_j^{(i)} > 0$ and

$$Y_{:,j} = \sum_{\ell>j} \frac{\alpha_\ell^{(j)} + \sum_{i<j} \alpha_i^{(j)} \beta_\ell^{(i)}}{1 - \sum_{i<j} \alpha_i^{(j)} \beta_j^{(i)}} Y_{:,\ell} \in \operatorname{range}_{\geq 0}(Y_{:,>j})$$

and $Y_{:,<j} \subseteq \operatorname{range}_{\geq 0}(Y_{:,\geq j}) = \operatorname{range}_{\geq 0}(Y_{:,>j})$. By induction, we conclude that $\operatorname{range}_{\geq 0}(Y_{:,\neq\mathcal{J}}) = \operatorname{range}_{\geq 0}(Y)$, i.e., all non-extreme columns are not needed to generate the non-negative range. Thus, $\mathcal{I} \neq \emptyset$ is the unique index set of all extreme columns with $\operatorname{range}_{\geq 0}(Y_{:,\mathcal{I}}) = \operatorname{range}_{\geq 0}(Y)$.

(i) Let Y contain at least one column which is a multiple of another one. Then one can prove that all non-extreme columns which are not a multiple of another one are not needed to generate the non-negative range by repeating the proof in part (iii) above for $\mathcal{J}$ denoting the corresponding index set. Then $\mathcal{I}$ can be processed as follows: We start with $\mathcal{I}$ being the index set of all extreme columns. If Y has columns $Y_{:,i_1} = \lambda_2 Y_{:,i_2} = \cdots = \lambda_\ell Y_{:,i_\ell}$, where $i_\mu \neq i_\nu$ for all $\mu \neq \nu$, $\lambda_\mu > 0$, and $Y_{:,i_1} \notin \operatorname{range}_{\geq 0}(Y_{:,\neq\{i_1,\ldots,i_\ell\}})$, then one adds exactly one index i_μ to $\mathcal{I}$. If Y has columns $Y_{:,i_1} = \lambda_2 Y_{:,i_2} = \cdots = \lambda_\ell Y_{:,i_\ell}$ with $i_\mu \neq i_\nu$ for all $\mu \neq \nu$, $\lambda_\mu \geq 0$, and $Y_{:,i_1} \in \operatorname{range}_{\geq 0}(Y_{:,\neq\{i_1,\ldots,i_\ell\}})$, then none of these indices is needed to generate $\operatorname{range}_{\geq 0}(Y)$. By construction of $\mathcal{I}$, directly $\operatorname{range}_{\geq 0}(Y) = \operatorname{range}_{\geq 0}(Y_{:,\mathcal{I}})$ and the minimality of $\mathcal{I}$ follows.

(ii) The fact that all columns of $Y_{:,\mathcal{I}}$ are extreme w.r.t. $Y_{:,\mathcal{I}}$ follows from the construction in parts (iii) and (i) above.

(iv) If Y has only extreme columns, $\mathcal{I} = [k]$ follows directly from part (iii).

Proof of Lemma 4.1. For $d = 2$, it holds true that

$$\begin{aligned}(x+y)^r &= \sum_{j=0}^{r} \binom{r}{j} x^{r-j} y^j = \sum_{j=0}^{r} \left(\sqrt{\binom{r}{r-j}} x^{r-j}\right) \left(\sqrt{\binom{r}{j}} y^j\right) \\ &= \left[\sqrt{\binom{r}{j}} x^j\right]_{j=r,\ldots,0}^{T} \left[\sqrt{\binom{r}{j}} y^j\right]_{j=0,\ldots,r}.\end{aligned}$$

Applying this recursively to $(\sum_{\mu=1}^{d-1} x_\mu + x_d)^r$ using

$$\begin{aligned}\sqrt{\binom{r}{\ell}}(x+y)^{r-\ell} &= \sum_{j=0}^{r-\ell} \sqrt{\binom{r}{\ell+j}} x^{r-\ell-j} \sqrt{\binom{r}{\ell}\binom{r-\ell}{j}^2 \binom{r}{\ell+j}^{-1}} y^j \\ &= \sum_{j=0}^{r-\ell} \sqrt{\binom{\ell}{\ell-j}} x^{\ell-j} \sqrt{\binom{r-\ell}{j}\binom{\ell+j}{\ell}} y^j\end{aligned}$$

with

$$\begin{aligned}\binom{r}{\ell}\binom{r-\ell}{j}^2 \binom{r}{\ell+j}^{-1} &= \frac{r!((r-\ell)!)^2(r-\ell-j)!}{\ell!(r-\ell)!(j!)^2((r-\ell-j)!)^2 r!} \\ &= \frac{(r-\ell)!}{j!(r-\ell-j)!} \frac{(\ell+j)!}{\ell!(\ell+j-\ell)!} = \binom{r-\ell}{j}\binom{\ell+j}{\ell}\end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

for $\ell \in [r]$ and $j \in \{0, \dots, r-\ell\}$ together with an index shift $j \leftarrow j+1$ leads to the above results.

Appendix D. Relaxation of the non-negativity constraint in Algorithm 3.1.

TABLE D.1
Geometric means of relative errors $(\|\mathbf{X}-\mathbf{P}\|_F)/\|\mathbf{P}\|_F$ *for* (4.1) *with* $\mathcal{X} = \{0, 1/n-1, \dots, 1\}$, $n \in \{2, 10\}$, *and* $r = 4$ *after* 25 *sweeps of Algorithm 2.2 or Algorithm A.1 with Algorithm 3.1 solving* (3.4) *for* $\varepsilon \in \{0, 10^{-16}, 10^{-12}, 10^{-8}, 10^{-4}, 1\}$ *in line 9 for* 30 *random initial values and geometric variances* $\sigma^2 \in [1.0, 6]$ *for* $\varepsilon \neq 1$ *and* $\sigma^2 \in [1.0, 61]$ *for* $\varepsilon = 1$.

n		ε	$d=20$	$d=40$	$d=60$	$d=80$	$d=100$	$d=120$
2	HER	0	$4.9\cdot10^{-4}$	$2.7\cdot10^{-3}$	$3.1\cdot10^{-3}$	$3.4\cdot10^{-3}$	$2.9\cdot10^{-3}$	$3.9\cdot10^{-3}$
		10^{-16}	$\mathbf{2.6\cdot10^{-4}}$	$2.6\cdot10^{-3}$	$3.4\cdot10^{-3}$	$3.8\cdot10^{-3}$	$3.2\cdot10^{-3}$	$4.1\cdot10^{-3}$
		10^{-12}	$5.4\cdot10^{-4}$	$\mathbf{2.2\cdot10^{-3}}$	$\mathbf{2.6\cdot10^{-3}}$	$\mathbf{2.7\cdot10^{-3}}$	$2.4\cdot10^{-3}$	$2.6\cdot10^{-3}$
		10^{-08}	$4.7\cdot10^{-4}$	$2.3\cdot10^{-3}$	$2.8\cdot10^{-3}$	$2.8\cdot10^{-3}$	$\mathbf{2.3\cdot10^{-3}}$	$\mathbf{2.4\cdot10^{-3}}$
		10^{-04}	$1.2\cdot10^{-2}$	$8.6\cdot10^{-3}$	$5.8\cdot10^{-3}$	$6.5\cdot10^{-3}$	$5.8\cdot10^{-3}$	$1.1\cdot10^{-2}$
		1	$4.2\cdot10^{-2}$	$4.3\cdot10^{-2}$	$4.3\cdot10^{-1}$	$3.4\cdot10^{-1}$	$3.8\cdot10^{-1}$	$3.5\cdot10^{-1}$
	ANLS	0	$4.2\cdot10^{-4}$	$2.8\cdot10^{-3}$	$3.4\cdot10^{-3}$	$3.8\cdot10^{-3}$	$3.2\cdot10^{-3}$	$3.5\cdot10^{-3}$
		10^{-16}	$\mathbf{3.2\cdot10^{-4}}$	$2.7\cdot10^{-3}$	$3.3\cdot10^{-3}$	$3.4\cdot10^{-3}$	$\mathbf{3.1\cdot10^{-3}}$	$2.8\cdot10^{-3}$
		10^{-12}	$6.1\cdot10^{-4}$	$3.0\cdot10^{-3}$	$3.0\cdot10^{-3}$	$3.5\cdot10^{-3}$	$3.2\cdot10^{-3}$	$3.5\cdot10^{-3}$
		10^{-08}	$5.2\cdot10^{-4}$	$\mathbf{2.4\cdot10^{-3}}$	$\mathbf{2.7\cdot10^{-3}}$	$\mathbf{2.9\cdot10^{-3}}$	$3.4\cdot10^{-3}$	$\mathbf{2.7\cdot10^{-3}}$
		10^{-04}	$1.5\cdot10^{-2}$	$9.5\cdot10^{-3}$	$7.6\cdot10^{-3}$	$6.4\cdot10^{-3}$	$7.5\cdot10^{-3}$	$7.0\cdot10^{-2}$
		1	$1.1\cdot10^{-2}$	$1.0\cdot10^{-2}$	$6.8\cdot10^{-1}$	$6.8\cdot10^{-1}$	$6.2\cdot10^{-1}$	$4.0\cdot10^{-1}$
10	HER	0	$1.9\cdot10^{-3}$	$1.6\cdot10^{-3}$	$2.4\cdot10^{-3}$	$\mathbf{3.1\cdot10^{-3}}$	$\mathbf{4.5\cdot10^{-3}}$	$5.5\cdot10^{-3}$
		10^{-16}	$2.2\cdot10^{-3}$	$1.7\cdot10^{-3}$	$\mathbf{2.3\cdot10^{-3}}$	$4.2\cdot10^{-3}$	$5.2\cdot10^{-3}$	$\mathbf{5.1\cdot10^{-3}}$
		10^{-12}	$\mathbf{1.7\cdot10^{-3}}$	$\mathbf{1.4\cdot10^{-3}}$	$2.4\cdot10^{-3}$	$3.8\cdot10^{-3}$	$\mathbf{4.5\cdot10^{-3}}$	$5.4\cdot10^{-3}$
		10^{-08}	$2.1\cdot10^{-3}$	$1.4\cdot10^{-3}$	$2.8\cdot10^{-3}$	$\mathbf{3.1\cdot10^{-3}}$	$5.3\cdot10^{-3}$	$5.8\cdot10^{-3}$
		10^{-04}	$3.1\cdot10^{-3}$	$6.8\cdot10^{-3}$	$1.7\cdot10^{-2}$	$1.9\cdot10^{-2}$	$3.1\cdot10^{-2}$	$3.6\cdot10^{-2}$
		1	$2.9\cdot10^{-2}$	$1.5\cdot10^{-1}$	1.0	1.0	1.0	1.0
	ANLS	0	$2.0\cdot10^{-3}$	$1.7\cdot10^{-3}$	$1.7\cdot10^{-3}$	$2.8\cdot10^{-3}$	$3.4\cdot10^{-3}$	$3.3\cdot10^{-3}$
		10^{-16}	$1.9\cdot10^{-3}$	$1.4\cdot10^{-3}$	$1.8\cdot10^{-3}$	$3.0\cdot10^{-3}$	$\mathbf{3.0\cdot10^{-3}}$	$\mathbf{2.9\cdot10^{-3}}$
		10^{-12}	$\mathbf{1.8\cdot10^{-3}}$	$1.8\cdot10^{-3}$	$\mathbf{1.6\cdot10^{-3}}$	$3.6\cdot10^{-3}$	$3.7\cdot10^{-3}$	$3.6\cdot10^{-3}$
		10^{-08}	$2.1\cdot10^{-3}$	$\mathbf{1.2\cdot10^{-3}}$	$2.0\cdot10^{-3}$	$\mathbf{2.5\cdot10^{-3}}$	$4.6\cdot10^{-3}$	$4.7\cdot10^{-3}$
		10^{-04}	$3.6\cdot10^{-3}$	$2.8\cdot10^{-3}$	$5.2\cdot10^{-3}$	$8.0\cdot10^{-3}$	$9.8\cdot10^{-3}$	$8.2\cdot10^{-3}$
		1	$5.0\cdot10^{-3}$	$2.1\cdot10^{-2}$	1.0	1.0	$5.3\cdot10^{-2}$	$4.5\cdot10^{-1}$

REFERENCES

[1] D. F. ANDERSON, G. CRACIUN, AND T. G. KURTZ, *Product-form stationary distributions for deficiency zero chemical reaction networks*, Bull. Math. Biology, 72 (2010), pp. 1947–1970.
[2] M. H. BAILEY, C. TOKHEIM, E. PORTA-PARDO ET AL., *Comprehensive characterization of cancer driver genes and mutations*, Cell, 173 (2018), pp. 371–385.e18.
[3] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.
[4] R. H. CHAN, *Iterative methods for overflow queueing models. I*, Numer. Math., 51 (1987), pp. 143–180.
[5] A. CICHOCKI, R. ZDUNEK, A. H. PHAN, AND S.-I. AMARI, *Nonnegative Matrix and Tensor Factorizations*, Wiley, Chichester, 2009.
[6] O. DEBALS, M. VAN BAREL, AND L. DE LATHAUWER, *Nonnegative matrix factorization using nonnegative polynomial approximations*, IEEE Signal Process. Lett., 24 (2017), pp. 948–952.
[7] A. FALCO, W. HACKBUSCH, AND A. NOUY, *Geometric structures in tensor representations*, Preprint on

ETNA
Kent State University and
Johann Radon Institute (RICAM)

arXiv, 2015. https://arxiv.org/abs/1505.03027

[8] P. GEORG, L. GRASEDYCK, M. KLEVER, R. SCHILL, R. SPANG, AND T. WETTIG, *Low-rank tensor methods for Markov chains with applications to tumor progression models*, J. Math. Biol., 86 (2022). https://doi.org/10.1007/s00285-022-01846-9

[9] N. GILLIS, *Sparse and unique nonnegative matrix factorization through data preprocessing*, J. Machine Learning Res., 13 (2012), pp. 3349–3386.

[10] ———, *Nonnegative Matrix Factorization*, SIAM, Philadelphia, 2021.

[11] L. GRASEDYCK, *Hierarchical singular value decomposition of tensors*, SIAM J. Matrix Anal. Appl., 31 (2009/10), pp. 2029–2054.

[12] L. GRASEDYCK, D. KRESSNER, AND C. TOBLER, *A literature survey of low-rank tensor approximation techniques*, GAMM-Mitt., 36 (2013), pp. 53–78.

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

[14] C. HAUTECOEUR, L. DE LATHAUWER, N. GILLIS, AND F. GLINEUR, *Least-squares methods for nonnegative matrix factorization over rational functions*, IEEE Trans. Signal Process., 71 (2023), pp. 1712–1724.

[15] J.-B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms I*, Springer, Berlin, 1993.

[16] S. HOLTZ, T. ROHWEDDER, AND R. SCHNEIDER, *On manifolds of tensors of fixed TT-rank*, Numer. Math., 120 (2012), pp. 701–731.

[17] ———, *The alternating linear scheme for tensor optimization in the tensor-train format*, SIAM J. Sci. Comput., 34 (2012), pp. A683–A713.

[18] K. HUANG, N. D. SIDIROPOULOS, AND A. P. LIAVAS, *A flexible and efficient algorithmic framework for constrained matrix and tensor factorization*, IEEE Trans. Signal Process., 64 (2016), pp. 5052–5065.

[19] K. HUANG, N. D. SIDIROPOULOS, AND A. SWAMI, *Non-negative matrix factorization revisited: uniqueness and algorithm for symmetric decomposition*, IEEE Trans. Signal Process., 62 (2014), pp. 211–224.

[20] N. KARMARKAR, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[21] L. KAUFMAN, *Matrix methods for queueing problems*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 525–552.

[22] T. G. KOLDA AND B. W. BADER, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

[23] S. KRÄMER, *Button Plot*, MATLAB Central File Exchange, Jan. 2023.

[24] A. N. LANGVILLE AND W. J. STEWART, *The Kronecker product and stochastic automata networks*, J. Comput. Appl. Math., 167 (2004), pp. 429–447.

[25] H. LAURBERG, M. G. CHRISTENSEN, M. D. PLUMBLEY, L. K. HANSEN, AND S. H. JENSEN, *Theorems on positive data: on the uniqueness of NMF*, Comput. Intell. Neurosci., 2008 (2008), pp. 1–9.

[26] D. D. LEE AND H. S. SEUNG, *Learning the parts of objects by non-negative matrix factorization*, Nature, 401 (1999), pp. 788–791.

[27] E. LEVINE AND T. HWA, *Stochastic fluctuations in metabolic pathways*, Proc. Natl. Acad. Sci. USA, 104 (2007), pp. 9224–9229.

[28] L. LOCONTE, N. DI MAURO, R. PEHARZ, AND A. VERGARI, *How to turn your knowledge graph embeddings into generative models*, in Advances in Neural Information Processing Systems, Vol. 36, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds., Curran Associates, NewYork, 2023, pp. 77713–77744.

[29] A. MAN SHUN ANG, J. E. COHEN, N. GILLIS, AND L. THI KHANH HIEN, *Accelerating block coordinate descent for nonnegative tensor factorization*, Numer. Linear Algebra Appl., 28 (2021), Paper No. e2373, 23 pages.

[30] H. MINC, *Nonnegative Matrices*, 2nd ed,, Wiley, New York, 1988.

[31] D. MITCHELL, N. YE, AND H. DE STERCK, *Nesterov acceleration of alternating least squares for canonical tensor decomposition: momentum step size selection and restart mechanisms*, Numer. Linear Algebra Appl., 27 (2020), Paper No. e2297, 24 pages.

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

[33] ———, *oseledets/TT-Toolbox*, GitHub, Jan. 2023.

[34] I. V. OSELEDETS AND E. E. TYRTYSHNIKOV, *Breaking the curse of dimensionality, or how to use SVD in many dimensions*, SIAM J. Sci. Comput., 31 (2009), pp. 3744–3759.

[35] B. PLATEAU AND W. J. STEWART, *Stochastic automata networks*, in Computational Probability, W. K. Grassmann, ed., International Series in Operations Research & Management Science, Springer, Boston, 2000, pp. 113–151.

[36] E. ROBEVA AND A. SEIGAL, *Duality of graphical models and tensor networks*, Inf. Inference, 8 (2019), pp. 273–288.

[37] R. SCHILL, S. SOLBRIG, T. WETTIG, AND R. SPANG, *Modelling cancer progression using mutual hazard networks*, Bioinformatics, 36 (2019), pp. 241–249.

[38] E. M. SHCHERBAKOVA, S. A. MATVEEV, A. P. SMIRNOV, AND E. E. TYRTYSHNIKOV, *Study of performance*

ETNA
Kent State University and
Johann Radon Institute (RICAM)

of low-rank nonnegative tensor factorization methods, Russian J. Numer. Anal. Math. Modelling, 38 (2023), pp. 231–239.

[39] G.-J. SONG AND M. K. NG, *Nonnegative low rank matrix approximation for nonnegative matrices*, Appl. Math. Lett., 105 (2020), Paper No. 106300, 7 pages.

[40] A. SULTONOV, S. MATVEEV, AND S. BUDZINSKIY, *Low-rank nonnegative tensor approximation via alternating projections and sketching*, Comp. Appl. Math., 42 (2023), Paper No. 68, 20 pages.

[41] THE MATHWORKS INC, *MATLAB, version 9.14.0 (R2023a)*, Natick.

[42] O. TAUSSKY, *A recurring theorem on determinants*, Amer. Math. Monthly, 56 (1949), pp. 672–676.

[43] A. USCHMAJEW AND B. VANDEREYCKEN, *The geometry of algorithms using hierarchical tensors*, Linear Algebra Appl., 439 (2013), pp. 133–166.

[44] C. F. VAN LOAN, *Tensor network computations in quantum chemistry*, Preprint, Cornell University, New York, 2008.

[45] N. VERVLIET, A. THEMELIS, P. PATRINOS, AND L. D. LATHAUWER, *A quadratically convergent proximal algorithm for nonnegative tensor decomposition*, in 28th European Signal Processing Conference, Amsterdam, IEEE Conference Proceedings, Los Alamitos, 2021, pp. 1020–1024.

[46] S. R. WHITE, *Density matrix formulation for quantum renormalization groups*, Phys. Rev. Lett., 69 (1992), pp. 2863–2866.

[47] Y. XU AND W. YIN, *A globally convergent algorithm for nonconvex optimization based on block coordinate update*, J. Sci. Comput., 72 (2017), pp. 700–734.

[48] Y. YE, *On the complexity of approximating a KKT point of quadratic programming*, Math. Programming, 80 (1998), pp. 195–211.

[49] Y. YE AND E. TSE, *An extension of Karmarkar's projective algorithm for convex quadratic programming*, Math. Programming, 44 (1989), pp. 157–179.

[50] R. ZDUNEK, *Alternating direction method for approximating smooth feature vectors in nonnegative matrix factorization*, in 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE Conference Proceedings, Los Alamitos, 2014, pp. 1–6.

[51] G. ZHOU, A. CICHOCKI, AND S. XIE, *Fast nonnegative matrix/tensor factorization based on low-rank approximation*, IEEE Trans. Signal Process. 60 (2012), pp. 2928–2940.

[52] G. M. ZIEGLER, *Lectures on Polytopes*, Springer, New York, 1995.

[53] S. ÖSTLUND AND S. ROMMER, *Thermodynamic limit of density matrix renormalization*, Phys. Rev. Lett., 75 (1995), pp. 3537–3540.