Non-negative Tensor Low-rank Decompositions Through the Lens of Information Geometry

Kazu Ghalamkari[1], **Jesper Løve Hinrich**[2], **Morten Mørup**[2]

[1]RIKEN AIP, [2]Technical University of Denmark
kazu.ghalamkari@riken.jp, jehi@dtu.dk, mmor@dtu.dk

Abstract

Non-negative tensor low-rank decompositions based on the Kullback–Leibler (KL) divergence minimization form non-convex optimization problems with the associated instability being a longstanding issue. This study introduces an information geometric analysis of such decompositions in order to enhance their stability. The key idea behind our analysis is to consider the tensor ranks as hidden variables and employ the EM algorithm and its information geometric view. We reveal that the instability in tensor decomposition arises from hidden variables breaking the flatness of the model manifold — the set of low-rank tensors. Consequently, we reformulate the tensor low-rank decomposition as iterative projections onto a flat model manifold of tensors without hidden variables, i.e., a set of rank-1 tensors, in higher-order tensor space than the original given tensor. This analysis bridges information geometry and tensor decomposition, resulting in a novel algorithm that ensures a monotonic decrease in the KL divergence regardless of the low-rank structure for which we presently consider the CP, Tucker, and Tensor Train decompositions.

1 Introduction

A tensor is a versatile data structure that describes multi-dimensional data and is widely used in computer vision (Panagakis et al. 2021), signal processing (Sidiropoulos et al. 2017), and data mining (Mørup 2011). Tensor low-rank decomposition, which approximates a given tensor with a linear combination of a small number of bases, enables a variety of applications such as density estimation (Novikov, Panov, and Oseledets 2021) and regression (Hendrikx et al. 2019). In many applications where the data are nonnegative, nonnegative tensor decompositions are often used to impose the non-negativity on the decomposed representation (Shashua and Hazan 2005) while promoting parts-based representations (Lee and Seung 1999). Tensor factorization requires optimizing the error between the input tensor and the reconstructed low-rank tensor. Although various error functions are developed, the KL divergence error is often used because it leads to improved robustness to outliers and noise when compared to the least square error (Gao et al. 2019).

However, both the KL-divergence and least-squares-based low-rank decompositions require non-convex optimization in most cases (Hillar and Lim 2013), and instability of the algorithms for such decompositions has been an issue in the tensor community. Existing multiplicative update (MU) methods (Kim, Cichocki, and Choi 2008) require careful discussion of convergence, and the batch gradient methods require tuning of the learning rate and batch size, and do not guarantee monotonically decreasing of the KL divergence (Glasser et al. 2019).

Recently, novel tensor modelings that avoid instability have been established (Ghalamkari and Sugiyama 2021, 2023) by understanding nonnegative tensor decompositions via information geometry (Amari 2016), which is the geometry of probability distributions. Information geometry introduces the concepts of flatness in the coordinate system of probability distributions, which is useful in order to formulate a problem as a convex optimization. Since normalized nonnegative tensors naturally correspond to joint discrete distributions (Vora, Gurumoorthy, and Rajwade 2021), nonnegative tensor decompositions can be described by information geometry. By formulating the decomposition so that the model manifold, i.e., the set of decomposed tensors, is flat, the Legendre decomposition (Sugiyama, Nakahara, and Tsuda 2018) and many-body decompositions (Ghalamkari, Sugiyama, and Kawahara 2023) have been developed which stably obtain the globally optimal solution by convex optimization. These decompositions are essentially different from conventional low-rank decompositions as they do not impose any low-rank structure on the tensor.

In contrast to these geometric-based tensor models, the geometry of traditional low-rank decompositions remains uncharted. In this study, we attempt to stabilize the nonnegative tensor low-rank decomposition by describing its information geometry. Specifically, we focus on the fact that the non-flatness of the low-rank manifold, i.e., the set of low-rank tensors, leads to difficulties in the optimization, and reformulate the tensor low-rank decomposition as stable iterative projections between flat manifolds within a higher-order tensor space. Interestingly, this projection can be viewed as a generalized rank-1 approximation of a higher-order tensor, and we prove that any rank-1 approximation optimizing the KL divergence is a convex problem regardless of the low-rank structure.

Our analysis provides a novel algorithm for tensor low-rank decompositions that monotonically reduces the KL di-

Copyright © 2025, AAAI 25 Workshop on Connecting Low-rank Representations in AI (https://april-tools.github.io/colorai/). All rights reserved.

vergence by iterative convex rank-1 approximations of higher-order tensors. In addition, given the closed-form formulas for the best rank-1 approximations, each projection in higher-order space can update all parameters simultaneously without the gradient method for popular low-rank decompositions such as Tucker (Kim and Choi 2007) and Tensor Train decomposition (Oseledets 2011), which provides a generalization of the existing Expectation-Maximization based CP decomposition (Huang and Sidiropoulos 2017; Yeredor and Haardt 2019).

Finally, we show experimentally that this algorithm performs competitively in terms of better solutions and speed of convergence compared to the conventional multiplicative updates and batched gradient approaches.

2 Preliminaries

2.1 Information Geometry

Information geometry is the geometry of the space in which each point is a probability distribution. In the following, we introduce information geometry for discrete distributions for our analysis of non-negative low-rank tensor decompositions.

Flatness and projections in information geometry Let $S(D)$ be the set of the entire discrete probability distribution $p(\boldsymbol{\omega})$ with D discrete random variables $\boldsymbol{\omega} = (\omega_1, \dots, \omega_D) \in \Omega$ for sample space Ω. In Euclidean space, the geodesic, i.e., the path to minimize the distance, is a straight line. In the space $S(D)$, two geodesics can be introduced: e-geodesics and m-geodesics. For two points, $p, q \in S(D)$, e- and m-geodesics are defined as

$$\{\, r_t \mid \log r_t = t \log p + (1-t) \log q + \phi(t) \,\},$$
$$\{\, r_t \mid r_t = tp + (1-t)q \,\},$$

respectively, where $0 \le t \le 1$ and $\phi(t)$ is a normalizing factor to make r_t a distribution. Let $\boldsymbol{\mathcal{U}}$ be a manifold in $S(D)$. We say that $\boldsymbol{\mathcal{U}}$ is an $e(m)$-flat manifold if the $e(m)$-geodesics between any two points in $\boldsymbol{\mathcal{U}}$ are included in $\boldsymbol{\mathcal{U}}$. For a given distribution p (q), the projection to find $q(p) \in \boldsymbol{\mathcal{U}}$ minimizing the KL-divergence $D(p, q)$ is called $m(e)$-projection from p onto $\boldsymbol{\mathcal{U}}$. The $m(e)$-projection onto an $e(m)$-flat manifold is in general a convex optimization problem (Amari 2021).

em-algorithm Let $\boldsymbol{\mathcal{M}}$ be a e-flat and $\boldsymbol{\mathcal{D}}$ be a m-flat manifold in $S(D)$. We find the nearest two points on $\boldsymbol{\mathcal{M}}$ and $\boldsymbol{\mathcal{D}}$ by iterating the m-projection from $q \in \boldsymbol{\mathcal{D}}$ onto $\boldsymbol{\mathcal{M}}$ and the e-projection from point $r \in \boldsymbol{\mathcal{M}}$ onto $\boldsymbol{\mathcal{D}}$. This procedure is called the em-algorithm. More specifically, given an initial point of $q_1 \in \boldsymbol{\mathcal{D}}$, the em-algorithm iteratively updates distributions by the following m- and e-step, respectively,

$$r_n = \operatorname*{arg\,min}_{r \in \boldsymbol{\mathcal{M}}} D(q_n, r), \quad q_{n+1} = \operatorname*{arg\,min}_{q \in \boldsymbol{\mathcal{D}}} D(q, r_n), \quad (1)$$

where the iteration $n = 1, 2, \dots$ continues until convergence. Although the flatness of $\boldsymbol{\mathcal{D}}$ and $\boldsymbol{\mathcal{M}}$ makes the solution of each step unique, the convergence point (q_∞, r_∞) of the algorithm depends on the initial value q_1. The expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977), a well-known method for maximum likelihood estimation with hidden variables, is a particular case of the em-algorithm (Hino, Akaho, and Murata 2024).

2.2 Non-negative tensor decomposition

Non-negative tensors as discrete distributions The normalized nonnegative tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times \cdots \times I_D}$ can be regarded as a joint discrete distribution whose index set is the sample space. More specifically, the element $\mathcal{T}_{i_1,\dots,i_D}$ corresponds to a value of the distribution $p(x_1 = i_1, \dots, x_D = i_D)$ whose sample space is the index set $\Omega_I = [I_1] \times \cdots \times [I_D]$. Thus, the tensor $\mathcal{T}$ is a point in $S(D)$. In the following, the indices of the tensor $\mathcal{T}$ are written as $\boldsymbol{i} = (i_1, \dots, i_D) \in \Omega_I$ for $i_d \in [I_d]$.

Tensor low-rank approximation We can extract features from tensor-formatted data by approximating the tensor with a low-rank structure. There are multiple possible choices to assume for the low-rank structure of the data. Although we address three typical low-rank structures, CP, Tucker, and Tensor Train decomposition, the generalization of the following discussion to any low-rank structure is straightforward. For $k \in \{\mathrm{CP}, \mathrm{Tucker}, \mathrm{Train}\}$ and a given D-th order tensor $\mathcal{T}$, the approximation by a low-rank structure can be formulated as,

$$\mathcal{T}_{\boldsymbol{i}} \simeq \mathcal{P}^k_{\boldsymbol{i}} = \sum_{\boldsymbol{r}} \mathcal{R}^k_{\boldsymbol{ir}}. \quad (2)$$

Here, the $(D + V^k)$-th order tensor $\mathcal{R}^k$ is defined as,

$$\mathcal{R}^{\mathrm{CP}}_{\boldsymbol{ir}} = \prod_{d=1}^{D} A^{(d)}_{i_d r}, \quad (3)$$

$$\mathcal{R}^{\mathrm{Tucker}}_{\boldsymbol{ir}} = \mathcal{G}_{\boldsymbol{r}} \prod_{d=1}^{D} A^{(d)}_{i_d r_d}, \quad (4)$$

$$\mathcal{R}^{\mathrm{Train}}_{\boldsymbol{ir}} = \prod_{d=1}^{D} \mathcal{G}^{(d)}_{r_{d-1} i_d r_d}, \quad (5)$$

for $V^{\mathrm{CP}} = 1$, $V^{\mathrm{Tucker}} = D$, $V^{\mathrm{Train}} = D - 1$, and $\boldsymbol{r} = (r_1, \dots, r_{V^k})$. Let Ω_R be the set of indices $\boldsymbol{r}$. That is, $\boldsymbol{r} \in \Omega_R \equiv [R_1] \times \cdots \times [R_{V^k}]$, where the degree of freedom of the index $(R_1, \dots, R_{V^k})$ is a hyperparameter called CP, Tucker, and train rank of the tensor $\mathcal{P}^k$, respectively, according to k. We assume that the tensor $\mathcal{T}$ is a normalized nonnegative tensor and discuss the approximation in terms of the KL divergence. Since the total sum of tensors is invariant in the KL divergence optimization (Ho and Van Dooren 2008), it holds that $\sum_{\boldsymbol{ir}} \mathcal{R}^k_{\boldsymbol{ir}} = 1$. Thus the tensor $\mathcal{R}^k$ is also identical to a distribution in $S(D + V^k)$. More specifically, $\mathcal{R}^k_{i_1,\dots,i_D,r_1,\dots,r_{V^k}}$ corresponds to

$$q(x_1 = i_1, \dots, x_D = i_D, h_1 = r_1, \dots, h_{V^k} = r_{V^k}).$$

Since the low-rank tensor $\mathcal{P}^k$ is a marginalization of the random variables $h_1, \dots, h_{V^k}$, we can regard the indices $\boldsymbol{r}$ as a hidden variable.

As a result, the tensor low-rank decomposition is a maximum likelihood estimation of distribution $\mathcal{T}$ with a model that has V^k hidden variables. We define the *low-rank manifold* as a set of low-rank tensors:

$$\boldsymbol{\mathcal{B}}^k = \left\{\, \mathcal{P}^k \mid \mathcal{P}^k_{\boldsymbol{i}} = \sum\nolimits_{\boldsymbol{r}} \mathcal{R}^k_{\boldsymbol{ir}} \,\right\},$$

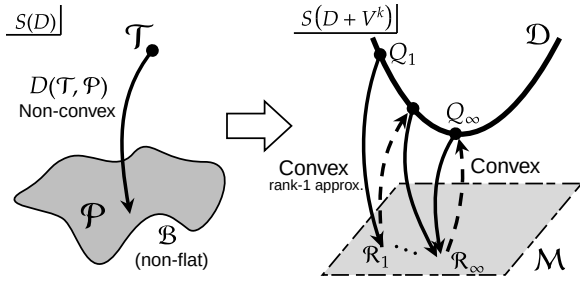

Figure 1: The tensor low-rank decomposition (left) can be viewed as the em-algorithm between the e-flat rank-1 manifold $\boldsymbol{\mathcal{M}}$ and the m-flat manifold $\boldsymbol{\mathcal{D}}$ in the higher dimensional space (right). The solid (dashed) arrows denote the $m(e)$-projection.

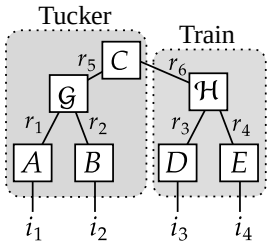

Figure 2: An example of a tensor tree structure represented by the tensor network diagram where the nodes represent factor tensors and edges connecting nodes represent mode products.

and the maximum likelihood estimation corresponds to the m-projection onto $\boldsymbol{\mathcal{B}}^k$. However, the manifold $\boldsymbol{\mathcal{B}}^k$ is not e-flat, and consequently the m-projection requires non-convex optimization, except for the rank-1 approximation, i.e., $(R_1, \ldots, R_{V^k}) = (1, \ldots, 1)$, where the m-projection becomes convex optimization as the sum of the hidden variables $\sum_{\boldsymbol{r}}$ disappears in the definition of the model manifold making it e-flat (Ghalamkari and Sugiyama 2022; Ghalamkari, Sugiyama, and Kawahara 2023). We call such a set of rank-1 tensors, i.e., tensors represented by a product without hidden variables, a *rank-1 manifold*. We also refer to the approximation of a given tensor by a tensor in a rank-1 manifold as *rank-1 approximation*.

3 Information geometry of tensor low-rank approximations

We want to find a low-rank tensor $\mathcal{P}^k$ that optimizes the KL divergence from the given tensor $\mathcal{T} \in S(D)$.

$$\mathcal{P}^k = \underset{\mathcal{P}^k \in \boldsymbol{\mathcal{B}}^k}{\arg\min}\, D(\mathcal{T}, \mathcal{P}^k) \tag{6}$$

However, the low-rank decomposition in $S(D)$ is unstable due to the non-convex optimization caused by hidden variables in the definition of the model manifold $\boldsymbol{\mathcal{B}}^k$. Therefore, we reformulate the problem to a rank-1 decomposition in the higher-order tensors space $S(D + V^k)$. Specifically, instead of a non-flat model manifold $\boldsymbol{\mathcal{B}}^k \subset S(D)$, we consider the rank-1 manifold $\boldsymbol{\mathcal{M}}^k = \{\mathcal{R} \mid \mathcal{R}_{\boldsymbol{ir}} = \mathcal{R}^k_{\boldsymbol{ir}}\}$, which is an e-flat manifold in $S(D + V^k)$ as shown in Proposition 1. Thus, the m-projection from any point in $S(D + V^k)$ to $\boldsymbol{\mathcal{M}}^k$ is a convex optimization problem as we show in Theorem 1. We also define the set of tensors $\boldsymbol{\mathcal{D}}$ that yields the given tensor $\mathcal{T}$ by marginalization, that is,

$$\boldsymbol{\mathcal{D}} = \left\{ \mathcal{Q} \mid \sum\nolimits_{\boldsymbol{r}} \mathcal{Q}_{\boldsymbol{ir}} = \mathcal{T}_{\boldsymbol{i}} \right\},$$

which is called the *data manifold*. The data manifold is m-flat in $S(D+V^k)$ as seen in Proposition 2. The tensor $\mathcal{T} \in S(D)$ corresponds to the manifold $\boldsymbol{\mathcal{D}} \subset S(D + V^k)$. As we see in Proposition 3, it holds that

$$D(\mathcal{T}, \mathcal{P}^k) \leq D(\mathcal{Q}, \mathcal{R})$$

for any tensor $\mathcal{Q} \in \boldsymbol{\mathcal{D}}$ and low-rank tensor $\mathcal{P}^k = \sum_{\boldsymbol{r}} \mathcal{R}_{\boldsymbol{ir}} \in \boldsymbol{\mathcal{M}}$. Instead of minimizing the objective function $D(\mathcal{T}, \mathcal{P}^k)$ directly, the em-algorithm between $\boldsymbol{\mathcal{D}}$ and $\boldsymbol{\mathcal{M}}^k$ minimizes $D(\mathcal{Q}, \mathcal{R})$ for $\mathcal{Q} \in \boldsymbol{\mathcal{D}}$ in e-step and $\mathcal{R} \in \boldsymbol{\mathcal{M}}^k$ in m-step in $S(D+V^k)$, iteratively. The e-flatness of data manifold $\boldsymbol{\mathcal{D}}$ and m-flatness of the rank-1 manifold $\boldsymbol{\mathcal{M}}$ guarantees the uniqueness and convexity of the e-step and m-step, respectively. Finally, we show that the em-algorithm between $\boldsymbol{\mathcal{D}}$ and $\boldsymbol{\mathcal{M}}^k$ decreases the objective function $D(\mathcal{T}, \mathcal{P}^k)$ monotonically in Theorem 2. Despite the convergence of the proposed algorithm, the solution depends on the initial value, and global convergence is not guaranteed.

We have now reduced the non-convex low-rank decomposition for a D-th order tensor to the iterative convex e-projection and m-projection in a $(D + V^k)$-th order tensor space as seen in Figure 1. We note that the CP decomposition based on the EM algorithm has been developed in (Huang and Sidiropoulos 2017). Our work generalizes the method for various low-rank structures inspired by the above information geometric analysis.

From here on, we discuss the specific procedures for these projections.

e-projection onto the data-manifold The Proposition 4 shows that the e-projection from $\mathcal{R} \in \boldsymbol{\mathcal{M}}$ onto $\boldsymbol{\mathcal{D}}$ is given as

$$\mathcal{Q}_{\boldsymbol{ir}} = \frac{\mathcal{T}_{\boldsymbol{i}} \mathcal{R}_{\boldsymbol{ir}}}{\sum_{\boldsymbol{r}} \mathcal{R}_{\boldsymbol{ir}}}. \tag{7}$$

m-projection onto the model-manifold This projection is a rank-1 approximation optimizing the KL divergence, which can be exactly solved by the natural gradient method without initial value dependency regardless of the low-rank structure on $\mathcal{P}$ (Ghalamkari, Sugiyama, and Kawahara 2023). In addition, when we impose a CP, Tucker, or Train structure on the tensor $\mathcal{P}$, the destination of each m-projection can be given by a closed-form solution. More specifically, for the CP structure defined in Equation (3), the destination of m-projection from $\mathcal{Q} \in \boldsymbol{\mathcal{D}}$ onto $\boldsymbol{\mathcal{M}}$ is given as:

$$A^{(d)}_{i_d r} = \frac{\sum_{\boldsymbol{i} \in \Omega_I^{\setminus d}} \mathcal{Q}_{\boldsymbol{i}r}}{\mu^{1/D} \left(\sum_{\boldsymbol{i} \in \Omega_I} \mathcal{Q}_{\boldsymbol{i}r} \right)^{1-1/D}}, \tag{8}$$

where $\mu = \sum_{\boldsymbol{i}\in\Omega_I}\sum_{r\in\Omega_R}\mathcal{Q}_{\boldsymbol{i}r}$, and for the Tucker structure defined in Equation (4), the destination is given as:

$$\mathcal{G}_{\boldsymbol{r}} = \frac{\sum_{\boldsymbol{i}\in\Omega_I}\mathcal{Q}_{\boldsymbol{i}\boldsymbol{r}}}{\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R}\mathcal{Q}_{\boldsymbol{i}\boldsymbol{r}}}, \quad A^{(d)}_{i_d r_d} = \frac{\sum_{\boldsymbol{i}\in\Omega_I^{\backslash d}}\sum_{\boldsymbol{r}\in\Omega_R^{\backslash d}}\mathcal{Q}_{\boldsymbol{i}\boldsymbol{r}}}{\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R^{\backslash d}}\mathcal{Q}_{\boldsymbol{i}\boldsymbol{r}}}, \tag{9}$$

and for the Train structure defined in Equation (5), the destination is given as:

$$\mathcal{G}^{(d)}_{r_{d-1}i_d r_d} = \frac{\sum_{\boldsymbol{i}\in\Omega_I^{\backslash d}}\sum_{\boldsymbol{r}\in\Omega_R^{\backslash d,d-1}}\mathcal{Q}_{\boldsymbol{i}\boldsymbol{r}}}{\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R^{\backslash d}}\mathcal{Q}_{\boldsymbol{i}\boldsymbol{r}}}, \tag{10}$$

where the symbol Ω with upper indices refers to the index set for all indices other than the upper indices, e.g.,

$$\Omega_I^{\backslash d} = [I_1]\times\cdots\times[I_{d-1}]\times[I_{d+1}]\times\cdots\times[I_D],$$
$$\Omega_R^{\backslash d,d-1} = [R_1]\times\cdots\times[R_{d-2}]\times[R_{d+1}]\times\cdots\times[R_V].$$

Please refer to Section A for these proofs. Inserting Equation (7) into Equations (8), (9), and (10) yields the simultaneous update rule for all parameters.

Moreover, we obtain the projection destination of the m-step in closed form for any tree low-rank structure (Liu, Long, and Zhu 2018) by combining Equations (8), (9), and (10). As an example of the tensor tree structure given in Figure 2, the objective function in the m-projection can be decoupled as

$$D(\mathcal{Q}^{\text{Tucker}}, \mathcal{R}^{\text{Tucker}}) + D(\mathcal{Q}^{\text{Train}}, \mathcal{R}^{\text{Train}})$$

where we define $\mathcal{Q}^{\text{Tucker}}_{i_1 i_2 r_1 r_2 r_5} = \sum_{i_3 i_4 r_3 r_4 r_6}\mathcal{Q}_{\boldsymbol{i}\boldsymbol{r}}$ and $\mathcal{Q}^{\text{Train}}_{i_3 i_4 r_3 r_4 r_6} = \sum_{i_1 i_2 r_1 r_2 r_5}\mathcal{Q}_{\boldsymbol{i}\boldsymbol{r}}$. We can optimize both decoupled terms by the closed-form solution given in Equations (9) and (10). We provide theoretical support for this procedure, including normalization conditions, in Section C.

We call the proposed algorithm em-NTF and summarize the procedure in Algorithm 1 in the Appendix.

4 Numerical Experiments

While our framework can be applied to general low-rank structures, we here numerically examine the effectiveness of the proposed em-Tucker and em-Train decomposition by comparing to KLNTDMU (Kim, Cichocki, and Choi 2008; Marmoret and Cohen 2020) and MPS (Glasser et al. 2019), respectively. The proposed em-Tucker(Train) and the baseline KLNTDMU(MPS) are the same model optimizing the same objective function with different methods. KLNTDMU is based on the multiplicative update rule, and MPS is based on the batch-gradient method. We used two datasets, DMFT (Simonoff 2003) and Hayesroth (Hayes-Roth and Hayes-Roth 1977), whose tensor sizes are $(9, 7, 2, 3, 6)$ and $(4, 4, 4, 4)$, respectively. We defined the $(R_1, \ldots, R_V)$ as $(2, \ldots, 2)$ for DMFT and $(5, \ldots, 5)$ for Hayesroth. For simplicity, we evaluated the negative log-likelihood $-\sum_{\boldsymbol{i}}\mathcal{T}_{\boldsymbol{i}}\log\mathcal{P}_{\boldsymbol{i}}$ instead of the KL divergence. We note that minimizing the negative likelihood is identical to minimizing the KL divergence.

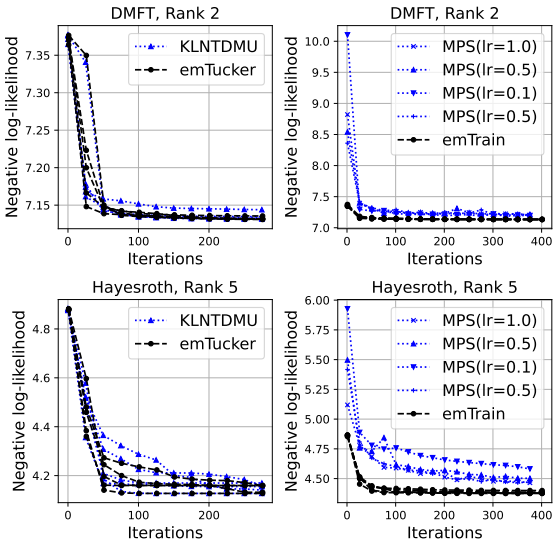

Figure 3: Experimental results for DMFT dataset (top), Hayesroth dataset (bottom).

Experiments for em-Tucker We repeated em-Tucker and KLNTDMU for two datasets five times with random initialization and plotted the negative log-likelihood at each iteration at the left of Figure 3. We verify that em-Tucker is able to find a comparable solution as fast as the baseline KLNTDMU.

Experiments for em-Train Since MPS requires learning rate tuning, we compared the objective function for each iteration while varying the learning rate. In contrast to the MPS, where monotonic decrease of the objective function is not guaranteed, we can see that the proposed em-Train monotonically decreases the error function and simultaneously updates all parameters in each iteration, resulting in finding better solutions, as seen in the right of Figure 3.

5 Conclusion

We revealed that non-negative tensor low-rank decomposition can be understood as maximum likelihood estimation with a model that has hidden variables corresponding to tensor ranks, and these hidden variables cause instability in the KL divergence minimization. To avoid this issue, inspired by information geometry, we reformulate non-negative tensor low-rank decomposition as iterative projections among flat manifolds, the set of rank-1 tensors and data manifold, in a higher order tensor space than the original tensor. Our analysis not only bridges information geometry and tensor low-rank decomposition but also forms a novel optimization framework, em-NTF. Although the proposed em-NTF is guaranteed to converge through iterative convex optimizations, i.e., e-step and m-step, the convergence point is not guaranteed to be a globally optimal solution. We also note that the higher-order tensor $\mathcal{Q} \in \boldsymbol{\mathcal{D}}$ is sparse if the tensor $\mathcal{T}$ is sparse, as we see in Equation (7), which makes our algorithm scalable.

Acknowledgements

This work was supported by RIKEN, Special Postdoctoral Researcher Program, ROIS, NII Open Collaborative Research 2025 Grant Number 24FP07, and JST, CREST Grant Number JPMJCR1913, Japan (GK) and the Novo Nordisk Foundation, Grant Number NNF23OC0083524 (MM).

References

Amari, S. 2016. *Information geometry and its applications*, volume 194. Springer.

Amari, S. 2021. Information geometry. *Japanese Journal of Mathematics*, 16(1): 1–48.

Chi, E. C.; and Kolda, T. G. 2012. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4): 1272–1299.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1): 1–22.

Ermiş, B.; Acar, E.; and Cemgil, A. T. 2015. Link prediction in heterogeneous data via generalized coupled tensor factorization. *Data Mining and Knowledge Discovery*, 29: 203–236.

Gao, L.; Dong, P.; Zheng, N.; and Tian, Y. 2019. Enhanced NMF Separation of Mixed Signals in Strong Noise Environment. *IEEE Access*, 7: 84649–84657.

Ghalamkari, K.; and Sugiyama, M. 2021. Fast Tucker Rank Reduction for Non-Negative Tensors Using Mean-Field Approximation. In *Advances in Neural Information Processing Systems*, volume 34, 443–454. Virtual Event.

Ghalamkari, K.; and Sugiyama, M. 2022. Fast Rank-1 NMF for Missing Data with KL Divergence. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, 2927–2940. Virtual Event.

Ghalamkari, K.; and Sugiyama, M. 2023. Non-negative Low-rank Approximations for Multi-dimensional Arrays on Statistical Manifold. *Information Geometry*, 6: 257–292.

Ghalamkari, K.; Sugiyama, M.; and Kawahara, Y. 2023. Many-body Approximation for Non-negative Tensors. In *Advances in Neural Information Processing Systems*, volume 36, 257–292. New Orleans, US.

Glasser, I.; Sweke, R.; Pancotti, N.; Eisert, J.; and Cirac, I. 2019. Expressive power of tensor-network factorizations for probabilistic modeling. *Advances in neural information processing systems*, 32.

Hayes-Roth, B.; and Hayes-Roth, F. 1977. Hayes-Roth. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5501T.

Hendrikx, S.; Boussé, M.; Vervliet, N.; and De Lathauwer, L. 2019. Algebraic and Optimization Based Algorithms for Multivariate Regression Using Symmetric Tensor Decomposition. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 475–479.

Hillar, C. J.; and Lim, L.-H. 2013. Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6): 1–39.

Hino, H.; Akaho, S.; and Murata, N. 2024. Geometry of EM and related iterative algorithms. *Information Geometry*, 7(Suppl 1): 39–77.

Ho, N.-D.; and Van Dooren, P. 2008. Non-negative matrix factorization with fixed row and column sums. *Linear Algebra and its Applications*, 429(5): 1020–1025.

Huang, K.; and Sidiropoulos, N. D. 2017. Kullback-Leibler principal component for tensors is not NP-hard. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, 693–697. IEEE.

Jensen, J. L. W. V. 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1): 175–193.

Kim, Y.-D.; and Choi, S. 2007. Nonnegative Tucker Decomposition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Kim, Y.-D.; Cichocki, A.; and Choi, S. 2008. Nonnegative Tucker decomposition with alpha-divergence. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1829–1832.

Krompaß, D.; Nickel, M.; Jiang, X.; and Tresp, V. 2013. Nonnegative tensor factorization with rescal. In *Tensor Methods for Machine Learning, ECML workshop*, 1–10.

Kırbız, S.; and Günsel, B. 2014. A multiresolution nonnegative tensor factorization approach for single channel sound source separation. *Signal Processing*, 105: 56–69.

Lee, D. D.; and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755): 788–791.

Liu, Y.; Long, Z.; and Zhu, C. 2018. Image completion using low tensor tree rank and total variation minimization. *IEEE Transactions on Multimedia*, 21(2): 338–350.

Marmoret, A.; and Cohen, J. 2020. nn_fac: Nonnegative Factorization techniques toolbox.

Mørup, M. 2011. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1): 24–40.

Novikov, G. S.; Panov, M. E.; and Oseledets, I. V. 2021. Tensor-train density estimation. In *Uncertainty in artificial intelligence*, 1321–1331. PMLR.

Oseledets, I. V. 2011. Tensor-Train Decomposition. *SIAM Journal on Scientific Computing*, 33(5): 2295–2317.

Panagakis, Y.; Kossaifi, J.; Chrysos, G. G.; Oldfield, J.; Nicolaou, M. A.; Anandkumar, A.; and Zafeiriou, S. 2021. Tensor Methods in Computer Vision and Deep Learning. *Proceedings of the IEEE*, 109(5): 863–890.

Phan, A. H.; and Cichocki, A. 2008. Fast and efficient algorithms for nonnegative Tucker decomposition. In *Advances in Neural Networks-ISNN 2008: 5th International Symposium on Neural Networks, ISNN 2008, Beijing, China, September 24-28, 2008, Proceedings, Part II 5*, 772–782. Springer.

Shashua, A.; and Hazan, T. 2005. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, 792–799.

Sidiropoulos, N. D.; De Lathauwer, L.; Fu, X.; Huang, K.; Papalexakis, E. E.; and Faloutsos, C. 2017. Tensor Decomposition for Signal Processing and Machine Learning. *IEEE Transactions on Signal Processing*, 65(13): 3551–3582.

Simonoff, J. S. 2003. *Analyzing categorical data*, volume 496. Springer.

Sugiyama, M.; Nakahara, H.; and Tsuda, K. 2018. Legendre Decomposition for Tensors. In *Advances in Neural Information Processing Systems 31*, 8825–8835. Montréal, Canada.

Takeuchi, K.; Tomioka, R.; Ishiguro, K.; Kimura, A.; and Sawada, H. 2013. Non-negative multiple tensor factorization. In *2013 IEEE 13th International Conference on Data Mining*, 1199–1204. IEEE.

Vora, J.; Gurumoorthy, K. S.; and Rajwade, A. 2021. Recovery of joint probability distribution from one-way marginals: Low rank tensors and random projections. In *2021 IEEE Statistical Signal Processing Workshop (SSP)*, 481–485. IEEE.

Yeredor, A.; and Haardt, M. 2019. Maximum likelihood estimation of a low-rank probability mass tensor from partial observations. *IEEE Signal Processing Letters*, 26(10): 1551–1555.

Zhao, Q.; Zhou, G.; Xie, S.; Zhang, L.; and Cichocki, A. 2016. Tensor ring decomposition. *arXiv preprint arXiv:1606.05535*.

Appendix

A Proofs

A.1 Dual flat structure in non-negative low-rank decomposition

We prove the flatness of the data manifold $\mathcal{D}$ and rank-1 manifold $\mathcal{M}$ in $S(D+V^k)$. In addition, based on the flatness of the rank-1 manifold, we show that any rank-1 approximation that optimizes the KL divergence is a convex optimization problem. Furthermore, we show the convergence of the proposed em-NTD in Theorem 2.

Proposition 1. *Regardless of low-rank structure, rank-1 tensor space is e-flat manifold.*

Proof. We consider K-th order tensors $\mathcal{R}$ that can be factorized with L factor tensors $\mathcal{Z}^1, \dots, \mathcal{Z}^L$ as

$$\mathcal{R}_{\boldsymbol{v}} = \prod_{l=1}^{L} \mathcal{Z}^l_{v(l)} \tag{11}$$

where $v(l)$ is a subset of indices $\boldsymbol{v} = (v_1, \dots, v_K)$. We can immediately confirm that the above equation becomes rank-1 CP, Tucker, and train decomposition when we define $v(l)$ appropriately. Since no sum for any indices appears in Equation (11), approximating a tensor in the form of Equation (11) is a rank-1 approximation. Thus, we define the set of rank-1 factorized tensors

$$\mathcal{M} = \left\{ \mathcal{R} \mid \mathcal{R}_{\boldsymbol{v}} = \prod_{l=1}^{L} \mathcal{Z}^l_{v(l)} \right\} \tag{12}$$

as a manifold in $S(K)$. For two tensors $\mathcal{R}^1, \mathcal{R}^2$ in the manifold $\mathcal{M}$, we consider its e-geodesics $\mathcal{U}$ satisfying

$$\log \mathcal{U}_{\boldsymbol{v}} = t \log \mathcal{R}^1_{\boldsymbol{v}} + (1-t) \log \mathcal{R}^2_{\boldsymbol{v}} + \phi(t)$$

where $\phi(t)$ is appropriately defined normalizing factor. We represent the above equation by factor tensors as

$$\begin{aligned} \log \mathcal{U}_{\boldsymbol{v}} &= t \log \prod_{l=1}^{L} \mathcal{Z}^{1,l}_{v(l)} + (1-t) \log \prod_{l=1}^{L} \mathcal{Z}^{2,l}_{v(l)} + \phi(t) \\ &= \log \prod_{l=1}^{L} (\mathcal{Z}^{1,l}_{v(l)})^t (\mathcal{Z}^{2,l}_{v(l)})^{1-t} + \phi(t) \end{aligned}$$

where we define $\mathcal{R}^m_{\boldsymbol{v}} = \prod_{l=1}^{L} \mathcal{Z}^{m,l}_{v(l)}$ for $m = 1, 2$. If we regard each term $(\mathcal{Z}^{1,l}_{v(l)})^t (\mathcal{Z}^{1,l}_{v(l)})^{1-t}$ as a factor, we can see the e-geodesics $\mathcal{U}$ belongs to the manifold $\mathcal{M}$. Thus, the manifold $\mathcal{M}$ is e-flat. □

Proposition 2. *For any tensor $\mathcal{T} \in S(D)$, the data manifold $\mathcal{D} = \{\mathcal{Q} \mid \sum_{\boldsymbol{r}} \mathcal{Q}_{\boldsymbol{ir}} = \mathcal{T}_{\boldsymbol{i}}\}$ is m-flat in $S(D+V)$.*

Proof. For any two tensors $\mathcal{Q}^1, \mathcal{Q}^2$ in the data manifold $\mathcal{D}$, we consider its m-geodesics,

$$\mathcal{U}_{\boldsymbol{ir}} = t\mathcal{Q}^1_{\boldsymbol{ir}} + (1-t)\mathcal{Q}^2_{\boldsymbol{ir}}.$$

Taking summing over indices $\boldsymbol{r}$, we can see the m-geodesics belong to the subspace $\mathcal{D}$ as follows:

$$\begin{aligned} \sum_{\boldsymbol{r}} \mathcal{U}_{\boldsymbol{ir}} &= t \sum_{\boldsymbol{r}} \mathcal{Q}^1_{\boldsymbol{ir}} + (1-t) \sum_{\boldsymbol{r}} \mathcal{Q}^2_{\boldsymbol{ir}} \\ &= t\mathcal{T}_{\boldsymbol{i}} + (1-t)\mathcal{T}_{\boldsymbol{i}} \\ &= \mathcal{T}_{\boldsymbol{i}}. \end{aligned}$$

Thus, $\mathcal{U}_{\boldsymbol{ir}} \in \mathcal{D}$ and the manifold $\mathcal{D}$ is m-flat. □

Theorem 1. *Any rank-1 non-negative decomposition optimizing the KL divergence is a convex optimization problem regardless of the low-rank structure.*

Proof. We consider factorization in the form of Equation (11). Specifically, for D-th order given non-negative tensor $\mathcal{T}$, we optimize the following problem as

$$\mathcal{R} = \arg\min_{\mathcal{R} \in \mathcal{M}} D(\mathcal{T}, \mathcal{R}) \tag{13}$$

where the model space $\mathcal{M}$ is introduced in Equation (12). Since the model manifold $\mathcal{M}$ is e-flat as shown in Proposition 1, the optimization in Equation (13) is a m-projection onto e-flat manifold, which is always a convex problem. Even if the tensor $\mathcal{T}$ is not normalized, it is straightforward to show that this optimization is still convex by the general relation $D(\lambda\mathcal{T}, \lambda\mathcal{R}) = \lambda D(\mathcal{T}, \mathcal{R})$ for any positive value λ. □

Proposition 3. *For given tensor $\mathcal{T} \in S(D)$ and a low-rank tensor $\mathcal{P} \in S(D)$ such that $\mathcal{P}_{\boldsymbol{i}} = \sum_{\boldsymbol{r} \in \Omega_R} \mathcal{R}_{\boldsymbol{ir}}$ where the tensor $\mathcal{R}$ is in a e-flat manifold $\mathcal{M} \subset S(D+V)$, the KL divergence from $\mathcal{T}$ to $\mathcal{P}$ satisfies following inequality*

$$D(\mathcal{T}, \mathcal{P}) \le D(\mathcal{Q}, \mathcal{R}) \tag{14}$$

for any tensor $\mathcal{Q} \in \mathcal{D} = \{\mathcal{Q} \mid \sum_{\boldsymbol{r}} \mathcal{Q}_{\boldsymbol{ir}} = \mathcal{T}_{\boldsymbol{i}}\} \subseteq S(D+V)$.

Proof.

$$\begin{aligned} D(\mathcal{T}, \mathcal{P}) &= \sum_{\boldsymbol{i} \in \Omega_I} \mathcal{T}_{\boldsymbol{i}} \log \frac{\mathcal{T}_{\boldsymbol{i}}}{\sum_{\boldsymbol{r}} \mathcal{R}_{\boldsymbol{ir}}} \\ &= \sum_{\boldsymbol{i} \in \Omega_I} \mathcal{T}_{\boldsymbol{i}} \log \mathcal{T}_{\boldsymbol{i}} - \sum_{\boldsymbol{i} \in \Omega_I} \mathcal{T}_{\boldsymbol{i}} \log \sum_{\boldsymbol{r} \in \Omega_R} \frac{\frac{\mathcal{Q}_{\boldsymbol{ir}}}{\mathcal{T}_{\boldsymbol{i}}} \mathcal{R}_{\boldsymbol{ir}}}{\frac{\mathcal{Q}_{\boldsymbol{ir}}}{\mathcal{T}_{\boldsymbol{i}}}} \\ &\le \sum_{\boldsymbol{i} \in \Omega_I} \mathcal{T}_{\boldsymbol{i}} \log \mathcal{T}_{\boldsymbol{i}} - \sum_{\boldsymbol{i} \in \Omega_I} \sum_{\boldsymbol{r} \in \Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \frac{\mathcal{T}_{\boldsymbol{i}} \mathcal{R}_{\boldsymbol{ir}}}{\mathcal{Q}_{\boldsymbol{ir}}} \\ &= \sum_{\boldsymbol{i} \in \Omega_I} \mathcal{T}_{\boldsymbol{i}} \log \mathcal{T}_{\boldsymbol{i}} - \sum_{\boldsymbol{i} \in \Omega_I} \sum_{\boldsymbol{r} \in \Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \frac{\mathcal{R}_{\boldsymbol{ir}}}{\mathcal{Q}_{\boldsymbol{ir}}} \\ &\qquad - \sum_{\boldsymbol{i} \in \Omega_I} \sum_{\boldsymbol{r} \in \Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \mathcal{T}_{\boldsymbol{i}} \\ &= \sum_{\boldsymbol{i} \in \Omega_I} \sum_{\boldsymbol{r} \in \Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \frac{\mathcal{Q}_{\boldsymbol{ir}}}{\mathcal{R}_{\boldsymbol{ir}}} \\ &= D(\mathcal{Q}, \mathcal{R}) \end{aligned}$$

where the following relation, called the Jensen inequality (Jensen 1906), is used:

$$f\left(\sum_{m=1}^{M} \lambda_m x_m \right) \le \sum_{m=1}^{M} \lambda_m f(x_m) \tag{15}$$

for any convex function $f : \mathbb{R} \to \mathbb{R}$ and real numbers $\lambda_1, \dots, \lambda_M$ that satisfies $\sum_{m=1}^{M} \lambda_m = 1$. □

Proposition 4. *The optimal e-projection from a tensor $\mathcal{R} \in S(D+V)$ onto the data manifold $\boldsymbol{\mathcal{D}} = \{\mathcal{Q} \mid \sum_{\boldsymbol{r}} \mathcal{Q}_{\boldsymbol{ir}} = \mathcal{T}_{\boldsymbol{i}}\} \subset S(D+V)$ is given as*

$$\mathcal{Q}_{\boldsymbol{ir}} = \frac{\mathcal{T}_{\boldsymbol{i}}\mathcal{R}_{\boldsymbol{ir}}}{\sum_{\boldsymbol{r}\in\Omega_R}\mathcal{R}_{\boldsymbol{ir}}}. \tag{16}$$

Proof. We prove this proposition by the fact that Equation (16) is the equality condition for Proposition 3. Specifically, we put Equation (16) into the KL divergence $D(\mathcal{T}, \mathcal{Q})$, then we obtain,

$$\begin{aligned} D(\mathcal{Q}, \mathcal{R}) &= \sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \frac{\mathcal{Q}_{\boldsymbol{ir}}}{\mathcal{R}_{\boldsymbol{ir}}} \\ &= \sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \frac{\mathcal{T}_{\boldsymbol{i}}\mathcal{R}_{\boldsymbol{ir}}}{\mathcal{P}_{\boldsymbol{i}}} \log \frac{\mathcal{T}_{\boldsymbol{i}}}{\mathcal{P}_{\boldsymbol{i}}} \\ &= \sum_{\boldsymbol{i}\in\Omega_I} \mathcal{T}_{\boldsymbol{i}} \log \frac{\mathcal{T}_{\boldsymbol{i}}}{\mathcal{P}_{\boldsymbol{i}}} = D(\mathcal{T}, \mathcal{P}). \end{aligned}$$

We used the relation $\sum_{\boldsymbol{r}} \mathcal{R}_{\boldsymbol{ir}} = \mathcal{P}_{\boldsymbol{i}}$. Since Jensen's inequality in Equation (15) shows

$$D(\mathcal{T}, \mathcal{P}) \leq D(\mathcal{Q}, \mathcal{R}), \tag{17}$$

the tensor $\mathcal{Q}$ in Equation (16) is optimal. □

Theorem 2. *For a given tensor $\mathcal{T}$ in $S(D)$, the em-algorithm between the data manifold $\boldsymbol{\mathcal{D}}$ and the model manifold $\boldsymbol{\mathcal{M}}$ monotonically decreases the KL divergence $D(\mathcal{T}, \mathcal{P})$ where the tensor $\mathcal{P}$ has low-rank structure such as $\mathcal{P}_{\boldsymbol{i}} = \sum_{\boldsymbol{r}} \mathcal{R}_{\boldsymbol{ir}}$.*

Proof. The e-step in iteration t updates $\mathcal{R}^{t-1}$ by optimal $\mathcal{R}^t$ to minimized the KL divergence $D(\mathcal{Q}, \mathcal{R})$ such as

$$D(\mathcal{T}, \mathcal{P}^t) = D(\mathcal{Q}^{t-1}, \mathcal{R}^t) \leq D(\mathcal{Q}^{t-1}, \mathcal{R}^{t-1})$$

where the low-rank tensor $\mathcal{P}^t$ can be written as $\mathcal{P}^t_{\boldsymbol{i}} = \sum_{\boldsymbol{r}} \mathcal{R}^t_{\boldsymbol{ir}}$. We used Proposition 3 for the left equal sign. The m-step in iteration t updates $\mathcal{R}^t$ by $\mathcal{R}^{t+1}$ to minimizes the KL divergence $D(\mathcal{Q}, \mathcal{R})$. Thus, it holds that

$$D(\mathcal{Q}^t, \mathcal{R}^t) \leq D(\mathcal{Q}^{t-1}, \mathcal{R}^t)$$

Again, the e-step in iteration $t+1$ updates $\mathcal{R}^t$ by optimal $\mathcal{R}^{t+1}$ to minimized the KL divergence $D(\mathcal{Q}, \mathcal{R})$ such as

$$D(\mathcal{T}, \mathcal{P}^{t+1}) = D(\mathcal{Q}^t, \mathcal{R}^{t+1}) \leq D(\mathcal{Q}^t, \mathcal{R}^t)$$

for the low-rank tensor $\mathcal{P}^{t+1}_{\boldsymbol{i}} = \sum_{\boldsymbol{r}} \mathcal{R}^{t+1}_{\boldsymbol{ir}}$. Combining the above three relations, we obtain

$$\begin{aligned} D(\mathcal{T}, \mathcal{P}^{t+1}) &= D(\mathcal{Q}^t, \mathcal{R}^{t+1}) \\ &\leq D(\mathcal{Q}^t, \mathcal{R}^t) \\ &\leq D(\mathcal{Q}^{t-1}, \mathcal{R}^t) = D(\mathcal{T}, \mathcal{P}^t) \end{aligned}$$

Thus, it holds that $D(\mathcal{T}, \mathcal{P}^{t+1}) \leq D(\mathcal{T}, \mathcal{P}^t)$. The algorithm converges, and the objective function monotonically decreases. □

A.2 Proofs for closed-form exact m-projections onto rank-1 manifolds

Theorem 3 (Optimal m-projection onto $\boldsymbol{\mathcal{M}}^{\mathrm{CP}}$ (Huang and Sidiropoulos 2017))**.** *For a given non-negative tensor $\mathcal{Q} \in \mathbb{R}^{I_1\times\cdots\times I_D\times R}_{\geq 0}$, its projection destination onto the e-flat manifold $\boldsymbol{\mathcal{M}} = \{\mathcal{R} \mid \mathcal{R}_{ir} = A_{i_1 r} \dots A_{i_D r}\}$ can be written as*

$$A^{(d)}_{i_d r} = \frac{\sum_{\boldsymbol{i}\in\Omega_I^{\setminus d}} \mathcal{Q}_{ir}}{\mu^{1/D}\left(\sum_{\boldsymbol{i}\in\Omega_I^{\setminus d}} \mathcal{Q}_{ir}\right)^{1-1/D}}, \quad \mu = \sum_{\boldsymbol{i}\in\Omega_I}\sum_{r\in\Omega_R} \mathcal{Q}_{ir}$$

Proof. Please refer to the original paper by Huang and Sidiropoulos (2017). □

Theorem 4 (The closed form of the optimal m-projection onto $\boldsymbol{\mathcal{M}}^{\mathrm{Tucker}}$)**.** *For a given non-negative tensor $\mathcal{Q} \in \mathbb{R}^{I_1\times\cdots\times I_D\times R_1\times\cdots\times R_D}_{\geq 0}$, its projection destination onto the e-flat manifold $\boldsymbol{\mathcal{M}} = \{\mathcal{R} \mid \mathcal{R}_{\boldsymbol{ir}} = \mathcal{G}_{\boldsymbol{r}} A_{i_1 r_1} \dots A_{i_D r_D}\}$ can be written as*

$$\mathcal{G}_{\boldsymbol{r}} = \frac{\sum_{\boldsymbol{i}\in\Omega_I} \mathcal{Q}_{\boldsymbol{ir}}}{\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}}}, \quad A^{(d)}_{i_d r_d} = \frac{\sum_{\boldsymbol{i}\in\Omega_I^{\setminus d}}\sum_{\boldsymbol{r}\in\Omega_{\boldsymbol{r}}^{\setminus d}} \mathcal{Q}_{\boldsymbol{ir}}}{\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R^{\setminus d}} \mathcal{Q}_{\boldsymbol{ir}}}.$$

Proof. This projection minimizes the KL divergence from the tensor $\mathcal{Q}$ onto the manifold $\boldsymbol{\mathcal{M}}$. Thus, the objective function can be written as

$$L(\mathcal{Q}; \mathcal{R}^{\mathrm{Tucker}}) = \sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \mathcal{R}^{\mathrm{Tucker}}_{\boldsymbol{ir}} \tag{18}$$

where

$$\mathcal{R}^{\mathrm{Tucker}}_{i_1\dots i_D r_1 \dots r_D} = \mathcal{G}_{r_1\dots r_D} A^{(1)}_{i_1 r_1} \dots A^{(D)}_{i_D r_D}.$$

We optimize the above objective function with normalizing condition $\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{R}^{\mathrm{Tucker}}_{\boldsymbol{ir}} = 1$. Then, we consider the following Lagrange function:

$$\begin{aligned} \mathcal{L} = &\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \mathcal{G}_{\boldsymbol{r}} A^{(1)}_{i_1 r_1} \dots A^{(D)}_{i_D r_D} \\ &- \lambda\left(\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{G}_{\boldsymbol{r}} A^{(1)}_{i_1 r_1} \dots A^{(D)}_{i_D r_D} - 1\right) \end{aligned}$$

To decouple the normalizing condition, we introduce scaled factor matrices $\tilde{A}^{(d)}$ as

$$\tilde{A}^{(d)}_{i_d r_d} = \frac{A^{(d)}_{i_d r_d}}{a^{(d)}_{r_d}}, \quad \text{where} \quad a^{(d)}_{r_d} = \sum_{i_d} A_{i_d r_d}, \tag{19}$$

and the scaled core tensor,

$$\tilde{\mathcal{G}}_{\boldsymbol{r}} = \mathcal{G}_{\boldsymbol{r}} a^{(1)}_{r_1} \dots a^{(D)}_{r_D}.$$

The normalizing condition $\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}^{\mathrm{Tucker}}_{\boldsymbol{ir}} = 1$ guarantees the normalization of the core tensor $\tilde{\mathcal{G}}$ as

$$\sum_{\boldsymbol{r}\in\Omega_R} \tilde{\mathcal{G}}_{\boldsymbol{r}} = 1. \tag{20}$$

The tensor $\mathcal{R}^{\mathrm{Tucker}}$ can be represented with the above introduced tensors as

$$\mathcal{R}^{\mathrm{Tucker}}_{\boldsymbol{ir}} = \mathcal{G}_{\boldsymbol{r}} A^{(1)}_{i_1 r_1} \dots A^{(D)}_{i_D r_D} = \tilde{\mathcal{G}}_{\boldsymbol{r}} \tilde{A}^{(1)}_{i_1 r_1} \dots \tilde{A}^{(D)}_{i_D r_D}.$$

We optimize $\tilde{\mathcal{G}}$ and $\tilde{A}^{(d)}_{i_d r_d}$ instead of $\mathcal{G}$ and $A^{(d)}_{i_d r_d}$. Thus the Lagrange function can be written as

$$\mathcal{L} = \sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \tilde{\mathcal{G}}_{\boldsymbol{r}} \tilde{A}^{(1)}_{i_1 r_1} \dots \tilde{A}^{(D)}_{i_D r_D} - \lambda \left(\sum_{\boldsymbol{r}} \tilde{\mathcal{G}}_{\boldsymbol{r}} - 1 \right) - \sum_{d=1}^{D} \sum_{r_d} \lambda^{(d)}_{r_d} \left(\sum_{i_d} \tilde{A}^{(d)}_{i_d r_d} - 1 \right). \tag{21}$$

The condition

$$\frac{\partial \mathcal{L}}{\partial \mathcal{G}_{\boldsymbol{r}}} = \frac{\partial \mathcal{L}}{\partial A^{(d)}_{i_d r_d}} = 0$$

leads equations

$$\tilde{\mathcal{G}}_{\boldsymbol{r}} = \frac{1}{\lambda} \sum_{\boldsymbol{i}\in\Omega_I} \mathcal{Q}_{\boldsymbol{ir}}, \quad \tilde{A}^{(d)}_{i_d r_d} = \frac{1}{\lambda^{(d)}_{r_d}} \sum_{\boldsymbol{i}\in\Omega_I^{\setminus d}} \sum_{\boldsymbol{r}\in\Omega_R^{\setminus d}} \mathcal{Q}_{\boldsymbol{ir}}.$$

The values of Lagrange multipliers are identified by the normalizing conditions (19) and (20) as

$$\lambda = \sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}}, \quad \lambda^{(d)}_{r_d} = \sum_{\boldsymbol{i}\in\Omega_I} \sum_{\boldsymbol{r}\in\Omega_R^{\setminus d}} \mathcal{Q}_{\boldsymbol{ir}}.$$

□

Theorem 5 (The closed form of the optimal m-projection onto $\boldsymbol{\mathcal{M}}^{\mathrm{Train}}$)**.** *For a given non-negative tensor $\mathcal{Q} \in \mathbb{R}^{I_1\times\cdots\times I_D \times R_1 \times \cdots \times R_{D-1}}_{\geq 0}$, its projection destination onto the e-flat manifold $\boldsymbol{\mathcal{M}} = \{\mathcal{R} \mid \mathcal{R}_{\boldsymbol{ir}} = \mathcal{G}^{(1)}_{i_1 r_1} \mathcal{G}^{(2)}_{r_1 i_2 r_2} \dots \mathcal{G}^{(D)}_{r_{D-1} i_D}\}$ can be written as*

$$\mathcal{G}^{(d)}_{r_{d-1} i_d r_d} = \frac{\sum_{\boldsymbol{i}\in\Omega_I^{\setminus d}} \sum_{\boldsymbol{r}\in\Omega_R^{\setminus d,d-1}} \mathcal{Q}_{\boldsymbol{ir}}}{\sum_{\boldsymbol{i}\in\Omega_I} \sum_{\boldsymbol{r}\in\Omega_R^{\setminus d}} \mathcal{Q}_{\boldsymbol{ir}}}$$

for $d = 1, \dots, D$, assuming $r_0 = r_D = 1$.

Proof. This projection minimizes the KL divergence from the tensor $\mathcal{Q}$ onto the manifold $\boldsymbol{\mathcal{M}}$. Thus, the objective function can be written as

$$L(\mathcal{Q}; \mathcal{R}^{\mathrm{Train}}) = \sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \mathcal{R}^{\mathrm{Train}}_{\boldsymbol{ir}}$$

where

$$\mathcal{R}^{\mathrm{Train}}_{i_1 \dots i_D r_1 \dots r_D} = \mathcal{G}^{(1)}_{i_1 r_1} \mathcal{G}^{(2)}_{r_1 i_2 r_2} \dots \mathcal{G}^{(D)}_{r_{D-1} i_D}.$$

We optimize the above objective function with normalizing condition $\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{R}^{\mathrm{Train}}_{\boldsymbol{ir}} = 1$. Then, we consider the following Lagrange function:

$$\mathcal{L} = \sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \mathcal{G}^{(1)}_{i_1 r_1} \mathcal{G}^{(2)}_{r_1 i_2 r_2} \dots \mathcal{G}^{(D)}_{r_{D-1} i_D} - \lambda \left(\sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{G}^{(1)}_{i_1 r_1} \mathcal{G}^{(2)}_{r_1 i_2 r_2} \dots \mathcal{G}^{(D)}_{r_{D-1} i_D} - 1 \right)$$

To decouple the normalizing condition, we introduce scaled core tensors $\tilde{\mathcal{G}}^{(1)}, \dots, \tilde{\mathcal{G}}^{(D-1)}$ that are normalized over r_{d-1} and i_d as

$$\tilde{\mathcal{G}}^{(d)}_{r_{d-1} i_d r_d} = \frac{g^{(d-1)}_{r_{d-1}}}{g^{(d)}_{r_d}} \mathcal{G}^{(d)}_{r_{d-1} i_d r_d},$$

where we define

$$g^{(d)}_{r_d} = \sum_{r_{d-1}} \sum_{i_d} \mathcal{G}^{(d)}_{r_{d-1} i_d r_d} g^{(d-1)}_{r_{d-1}},$$

with $g_{r_0} = 1$. We assume $r_0 = r_D = 1$. Using the scaled core tensors, the tensor $\mathcal{Q}^{\mathrm{Train}}$ can be written as

$$\begin{aligned} \mathcal{Q}^{\mathrm{Train}}_{i_1 \dots i_D r_1 \dots r_D} &= \mathcal{G}^{(1)}_{i_1 r_1} \mathcal{G}^{(2)}_{r_1 i_2 r_2} \dots \mathcal{G}^{(D)}_{r_{D-1} i_D} \\ &= \tilde{\mathcal{G}}^{(1)}_{i_1 r_1} \tilde{\mathcal{G}}^{(2)}_{r_1 i_2 r_2} \dots \tilde{\mathcal{G}}^{(D)}_{r_{D-1} i_D} \end{aligned}$$

with

$$\tilde{\mathcal{G}}^{(D)}_{r_{D-1} i_D} = \frac{1}{g^{(D-1)}_{r_{D-1}}} \mathcal{G}^{(D)}_{r_{D-1} i_D}. \tag{22}$$

The matrix $\tilde{\mathcal{G}}^{(D)}$ is normalized, satisfying $\sum_{r_{D-1}} \sum_{i_D} \tilde{\mathcal{G}}^{(D)}_{r_{D-1} i_D} = 1$. Thus, the Lagrange function can be written as

$$\begin{aligned} \mathcal{L} = \sum_{\boldsymbol{i}\in\Omega_I}\sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \tilde{\mathcal{G}}^{(1)}_{i_1 r_1} \tilde{\mathcal{G}}^{(2)}_{r_1 i_2 r_2} \dots \tilde{\mathcal{G}}^{(D)}_{r_{D-1} i_D} \\ - \sum_{d=1}^{D-1} \lambda^{(d)}_{r_d} \left(\sum_{r_{d-1}} \sum_{i_d} \tilde{\mathcal{G}}^{(d)}_{r_{d-1} i_d r_d} - 1 \right) \\ - \lambda^{(D)} \left(\sum_{r_{D-1}} \sum_{i_d} \tilde{\mathcal{G}}^{(D)}_{r_{D-1} i_D} - 1 \right). \end{aligned} \tag{23}$$

The critical condition

$$\frac{\partial \mathcal{L}}{\partial \mathcal{G}^{(d)}_{r_{d-1} i_d r_d}} = 0$$

leads the equation

$$\tilde{\mathcal{G}}^{(d)}_{r_{d-1} i_d r_d} = \frac{1}{\lambda^{(d)}_{r_d}} \sum_{\boldsymbol{i}\in\Omega_I^{\setminus d}} \sum_{\boldsymbol{r}\in\Omega_R^{\setminus d,d-1}} \mathcal{Q}_{\boldsymbol{ir}},$$

where the values of multipliers $\lambda^{(d)}$ are identified by the normalizing conditions in Equation (22) as

$$\lambda^{(d)}_{r_d} = \sum_{\boldsymbol{i}\in\Omega_I} \sum_{\boldsymbol{r}\in\Omega_R^{\setminus d}} \mathcal{Q}_{\boldsymbol{ir}}.$$

□

B Additonal remarks

B.1 Technical detail for tree low-rank structures

We here show how to decouple the m-projection for a tensor tree structure into solvable m-projections. In the following, we discuss the decomposition with low-rank structure as

Algorithm 1: em-NTF for CP, Tucker, or Train decomposition

input :Tensor $\mathcal{T} \in S(D)$, and tensor-rank $(R_1, \dots, R_V)$
Initialize $\mathcal{R} \in S(D+V)$;
repeat
  $\mathcal{P}_{\boldsymbol{i}} \leftarrow \sum_{\boldsymbol{r}} \mathcal{R}_{\boldsymbol{ir}}$;
  $\mathcal{Q}_{\boldsymbol{ir}} \leftarrow \mathcal{T}_{\boldsymbol{i}} \mathcal{R}_{\boldsymbol{ir}} / \sum_{\boldsymbol{r}} \mathcal{R}_{\boldsymbol{ir}}$; // e-step
  Update $\mathcal{R}$ using Equations (8), (9), or (10); // m-step
until *Convergence*;
return *Low-rank tensor* $\mathcal{P}$*;*

Algorithm 2: em-NTF for general low-rank decomposition

input :Tensor $\mathcal{T} \in S(D)$, and tensor-rank $(R_1, \dots, R_V)$
Initialize $\mathcal{R} \in S(D+V)$;
repeat
  $\mathcal{P}_{\boldsymbol{i}} \leftarrow \sum_{\boldsymbol{r}} \mathcal{R}_{\boldsymbol{ir}}$;
  $\mathcal{Q}_{\boldsymbol{ir}} \leftarrow \mathcal{T}_{\boldsymbol{i}} \mathcal{R}_{\boldsymbol{ir}} / \sum_{\boldsymbol{r}} \mathcal{R}_{\boldsymbol{ir}}$; // e-step
  Update $\mathcal{R}$ by many-body approximation (Ghalamkari, Sugiyama, and Kawahara 2023); // m-step
until *Convergence*;
return *Low-rank tensor* $\mathcal{P}$*;*

described in Figure 2 as an example, while the generalization to arbitrary tree low-rank structures is straightforward. When we decouple the m-projection, we need to guarantee that the normalizing condition

$$\sum_{\boldsymbol{i}\in\Omega_I} \sum_{\boldsymbol{r}\in\Omega_R} \mathcal{R}_{\boldsymbol{ir}} = 1 \tag{24}$$

is satisfied where we define

$$\mathcal{R}_{\boldsymbol{ir}} = \mathcal{G}_{r_1r_2r_5} A_{i_1r_1} B_{i_2r_2} C_{r_5r_6} D_{i_3r_3} \mathcal{H}_{r_3r_6r_4} E_{r_4i_4}. \tag{25}$$

We decouple the Lagrange function into independent parts.

More specifically, we define a single root tensor and introduce normalized factors that sums over the edges that lie below from the root. Although the choice of the root tensor is not unique, we let tensor $\mathcal{G}$ be the root tensor and introduce

$$\tilde{A}_{i_1r_1} = \frac{1}{a_{r_1}} A_{i_1r_1}, \quad \tilde{B}_{i_2r_2} = \frac{1}{b_{r_2}} B_{i_2r_2}, \tag{26}$$

$$\tilde{C}_{r_5r_6} = \frac{h_{r_6}}{c_{r_5}} C_{r_5r_6}, \quad \tilde{D}_{i_3r_3} = \frac{1}{d_{r_3}} D_{i_3r_3}, \tag{27}$$

$$\tilde{E}_{i_4r_4} = \frac{1}{e_{r_4}} E_{i_4r_4}, \quad \tilde{\mathcal{H}}_{r_3r_4r_6} = \frac{d_{r_3}e_{r_4}}{h_{r_6}} \mathcal{H}_{r_3r_4r_6} \tag{28}$$

where each normalizer is defined as

$$a_{r_1} = \sum_{i_1} A_{i_1r_1}, \quad b_{r_2} = \sum_{i_2} B_{i_2r_2},$$

$$c_{r_5} = \sum_{r_6} C_{r_5r_6} h_{r_6}, \quad d_{r_3} = \sum_{i_3} D_{i_3r_3},$$

$$e_{r_4} = \sum_{i_4} E_{i_4r_4}, \quad h_{r_6} = \sum_{r_3r_4} d_{r_3} e_{r_4} \mathcal{H}_{r_3r_4r_6},$$

then it holds that

$$\sum_{i_1} \tilde{A}_{i_1r_1} = \sum_{i_2} \tilde{B}_{i_2r_2} = \sum_{r_6} \tilde{C}_{r_5r_6}$$
$$= \sum_{i_3} \tilde{D}_{i_3r_3} = \sum_{i_4} \tilde{E}_{i_4r_4} = \sum_{r_3r_4} \tilde{\mathcal{H}}_{r_3r_4r_6} = 1.$$

We define the tensor $\tilde{\mathcal{G}}$ as $\tilde{\mathcal{G}}_{r_1r_2r_5} = a_{r_1} b_{r_2} c_{r_5} \mathcal{G}_{r_1r_2r_5}$ and putting Equations (26) and (27) into Equations (25) and (24), we obtain the normalizing condition for the root tensor $\mathcal{G}$ as

$$\sum_{r_1r_2r_5} \tilde{\mathcal{G}}_{r_1r_2r_5} = 1.$$

Then, the tensor $\mathcal{Q}$ can be written as

$$\mathcal{R}_{\boldsymbol{ir}} = \tilde{\mathcal{G}}_{r_1r_2r_5} \tilde{A}_{i_1r_1} \tilde{B}_{i_2r_2} \tilde{C}_{r_5r_6} \tilde{D}_{i_3r_3} \tilde{\mathcal{H}}_{r_3r_6r_4} \tilde{E}_{r_4i_4}. \tag{29}$$

The above approach to reduce scaling redundancy is illustrated in Figure 4. Finally, the original optimization problem with the Lagrange function

$$\mathcal{L} = \sum_{\boldsymbol{i}\in\Omega_I} \sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \mathcal{R}_{\boldsymbol{ir}} - \lambda \left(1 - \sum_{\boldsymbol{i}\in\Omega_I} \sum_{\boldsymbol{r}\in\Omega_R} \mathcal{R}_{\boldsymbol{ir}}\right)$$

is equivalent to the problem with the Lagrange function

$$\mathcal{L} = \mathcal{L}^{\text{Tucker}} + \mathcal{L}^{\text{Train}}$$

where

$$\mathcal{L}^{\text{Tucker}} = \sum_{\boldsymbol{i}\in\Omega_I} \sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \tilde{G}_{r_1r_2r_5} \tilde{A}_{i_1r_1} \tilde{B}_{i_2r_2} \tilde{C}_{r_5r_6}$$
$$+ \lambda^{\mathcal{G}} \left(\sum_{r_1r_2r_5} \tilde{\mathcal{G}}_{r_1r_2r_5} - 1\right) + \sum_{r_1} \lambda^A_{r_1} \left(\sum_{i_1} \tilde{A}_{i_1r_1} - 1\right)$$
$$+ \sum_{r_2} \lambda^B_{r_2} \left(\sum_{i_2} \tilde{B}_{i_2r_2} - 1\right) + \sum_{r_5} \lambda^C_{r_5} \left(\sum_{r_6} \tilde{C}_{r_5r_6} - 1\right),$$

which is equivalent to the Lagrange function for the Tucker

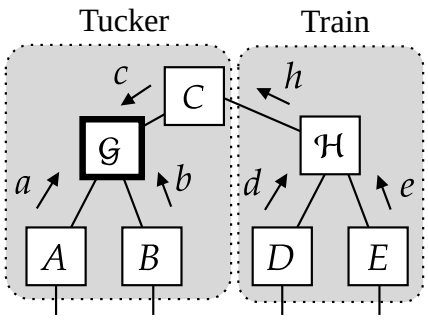

Figure 4: We normalize all tensors except for the root tensor, which is enclosed in a bold line. We then push the normalizer of each tensor, a, b, c, d, e, and h on the root tensor. The root tensor absorbs scaling redundancy. This procedure decouples the Lagrangian $\mathcal{L}$ into two independent problems, $\mathcal{L}^{\text{Tucker}}$ and $\mathcal{L}^{\text{Train}}$

decomposition given in Equation (21) and

$$\mathcal{L}^{\text{Train}} = \sum_{\boldsymbol{i}\in\Omega_I} \sum_{\boldsymbol{r}\in\Omega_R} \mathcal{Q}_{\boldsymbol{ir}} \log \tilde{\mathcal{H}}_{r_3 r_4 r_6} \tilde{D}_{i_3 r_3} \tilde{E}_{i_4 r_4} + \sum_{r_3} \lambda^D_{r_3} \left(\sum_{i_3} \tilde{D}_{i_3 r_3} - 1 \right) + \sum_{r_4} \lambda^E_{r_4} \left(\sum_{i_4} \tilde{E}_{i_4 r_4} - 1 \right) + \sum_{r_6} \lambda^{\mathcal{H}}_{r_6} \left(\sum_{r_3 r_4} \tilde{\mathcal{H}}_{r_3 r_4 r_6} - 1 \right),$$

which is also equivalent to the Lagrange function for the Train decomposition given in Equation (23) assuming $G^{(D)}$ is a normalized uniform tensor. For simplicity, we define tensors

$$\mathcal{Q}^{\text{Tucker}}_{i_1 i_2 r_1 r_2 r_5} = \sum_{i_3 i_4} \sum_{r_3 r_4 r_6} \mathcal{Q}_{\boldsymbol{ir}},$$

$$\mathcal{Q}^{\text{Train}}_{i_3 i_4 r_3 r_4 r_6} = \sum_{i_1 i_2} \sum_{r_1 r_2 r_5} \mathcal{Q}_{\boldsymbol{ir}},$$

then, solve these independent m-projections by the closed-form solution by Equations (9) and (10) for given tensors $\mathcal{Q}^{\text{Tucker}}$ and $\mathcal{Q}^{\text{Train}}$, respectively, and multiply solutions to get optimal tensor $\mathcal{Q}$ as Equation (29), which satisfied the normalizing condition in Equation (24).

B.2 em-NTF for general non-negative tensors

Non-negative tensor factorization optimizing the KL divergence is frequently used beyond density estimation and in various fields such as sound source separation (Kırbız and Günsel 2014), computer vision (Kim, Cichocki, and Choi 2008; Phan and Cichocki 2008), and data mining (Chi and Kolda 2012; Takeuchi et al. 2013; Krompaß et al. 2013; Ermiş, Acar, and Cemgil 2015). Although the given tensor $\mathcal{T}$ is not necessarily normalized in such applications, the proposed framework can be used for them as follows. First, we obtain the total sum of the input tensor $\mu = \sum_{\boldsymbol{i}} \mathcal{T}_{\boldsymbol{i}}$ and then perform the factorization on the normalized tensor $\mathcal{T}$ by dividing all elements of $\mathcal{T}$ by μ. Finally, all elements of the resulting tensor $\mathcal{P}$ are multiplied by μ. This procedure is justified by the property of the KL divergence, $D_{KL}(\mu\mathcal{P}, \mu\mathcal{T}) = \mu D_{KL}(\mathcal{P}, \mathcal{T})$, where μ is any positive value.

C Algorithms

We here provide the proposed algorithm. For CP, Tucker, and Tensor Train decomposition, we directly apply the closed form given in Equations (8), (9) and (10), which is summarized in Algorithm 1. For tensor tree decomposition, we decouple the m-step into solvable low-rank parts and update each part as we discussed in Sections 3 and B.1.

If we assume a low-rank structure that includes loops in the tensor network representation, such as tensor ring decomposition (Zhao et al. 2016), we perform the m-step by a natural gradient method since the closed-form update is no longer available as seen in Algorithm 2. In this case, the m-step is called a many-body approximation and still convex optimization. Please refer to the original paper by Ghalamkari, Sugiyama, and Kawahara (2023) for the procedure for many-body approximation.