Fast Tucker Rank Reduction for Non-Negative Tensors Using Mean-Field Approximation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present an efficient low-rank approximation algorithm for non-negative tensors. The algorithm is derived from our two findings: First, we show that rank-1 approximation for tensors can be viewed as a *mean-field approximation* by treating each tensor as a probability distribution. Second, we theoretically provide a sufficient condition for distribution parameters to reduce Tucker ranks of tensors and, interestingly, this sufficient condition can be achieved by iterative application of the mean-field approximation. Since the mean-field approximation is always given as a closed formula, our findings lead to a fast low-rank approximation algorithm without using a gradient method. We empirically demonstrate that our algorithm is faster than the existing non-negative Tucker rank reduction methods with achieving competitive or better approximation of given tensors.

1 Introduction

A multidimensional array, or a *tensor*, is a fundamental data structure in machine learning and statistical data analysis, and extraction of the essential information contained in tensors has been extensively studied [14, 20]. For second-order tensors; that is, matrices, *low-rank approximation* by singular value decomposition (SVD) is well established [13]. SVD always provides the best low-rank approximation in the sense of arbitrary unitarily invariant norms [29]. In contrast, the problem of low-rank approximation becomes much more challenging for tensors higher than the second order, where how to define the rank of tensors is even nontrivial. To date, two types of ranks, the CP-rank [17, 24] and the Tucker rank [9, 38], have been proposed, and low-rank approximation of tensors in terms of one of the above two ranks has been widely studied. Furthermore, non-negative low-rank approximation has been also developed for not only matrices such as NMF [26] but also tensors [27]. In particular, non-negative Tucker decomposition (NTD) [21] and its efficient variant lraSNTD [42] approximate a given non-negative tensor by a tensor with the lower Tucker rank.

While these approximations have been widely used in various domains such as image classification [23], recommendation [36], and denoising [12], efficient low-rank approximation is still fundamentally challenging. Even the simplest case, the rank-1 approximation in terms of minimizing the Least Squares (LS) error between a given tensor and a low-rank tensor, is known to be NP-hard [16]. Various methods have been developed to efficiently find approximate solutions in polynomial runtime [7, 8, 10, 25, 41]. If we use the Kullback–Leibler (KL) divergence instead of the LS error as a cost function, we can alleviate the problem as the best rank-1 approximation can be obtained in the closed formula [19]. However, the general case of low-rank approximation in terms of the KL divergence is also still under development.

In this paper, we present a fast low Tucker rank approximation method for non-negative tensors. To date, the majority of low-rank approximation methods are based on gradient decent using the derivative of the cost function, which often requires careful tuning of initialization and/or a tolerance

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

threshold. In contrast, our method is not based on a gradient method; the solution is directly obtained based on a closed formula, which we derive from information geometric treatment of tensors. Through an alternative parameterization of tensors by treating them as probability distributions in a statistical manifold, we theoretically provide a sufficient condition for such parameters, called the *bingo rule*, to reduce Tucker ranks of tensors. We then show that low-rank approximation is achieved by m*-projection*, which is one of the two canonical projections in information geometry [2], where a distribution (corresponding to a given non-negative tensor) is projected onto the subspace restricted by the bingo rule (corresponding to the set of non-negative low-rank tensors).

The key insight is that rank-1 approximation for non-negative tensors can be exactly solved by a *mean-field approximation*, a well-established method in physics that approximates a joint distribution by independent distributions [39], as we can represent any non-negative rank-1 tensor by a product of independent distributions. Moreover, we show that the bingo rule, our sufficient condition for tensor Tucker rank reduction, can be achieved by iterative applications of the mean-field approximation. Together with the fact that mean-field approximation is computed by m-projection in the closed form, we can derive our fast low Tucker rank approximation method without using a gradient method. Our theoretical analysis has a close relationship to [35]. Their proposal, called Legendre decomposition, also uses information geometric parameterization of tensors and solves the problem of tensor decomposition by a projection onto a subspace. Although we use the same information geometric formulation of tensors, they do not provide any connection to the Tucker ranks, and Tucker rank reduction is not guaranteed by their approach. Our method treats only non-negative tensors and cannot handle tensors with zero or negative values, which is the limitation of our method. Although experimental results show the usefulness of our method even for tensors including zeros, we always assume that every element of an input tensor is strictly positive in our theoretical discussion.

2 The Proposed Rank Reduction Algorithm

We propose a Tucker rank reduction algorithm, called *Legendre Tucker rank reduction* (LTR), which transforms a given tensor $\mathcal{P} \in \mathbb{R}^{I_1 \times \cdots \times I_d}$ into a tensor that approximates $\mathcal{P}$ with arbitrary reduced Tucker rank $(r_1, \ldots, r_d)$ specified by the user. The term "Legendre" of our algorithm comes from the fact that the theoretical support of our algorithm uses parameterization of tensors based on the *Legendre transformation*, which will be clarified in Section 3.

2.1 Problem setup and notation

First we define the Tucker rank of tensors and formulate the problem of non-negative Tucker rank reduction. The *Tucker rank* of a dth-order tensor $\mathcal{P} \in \mathbb{R}^{I_1 \times \cdots \times I_d}$ is defined as a tuple $(\mathrm{Rank}(\mathcal{P}^{(1)}), \ldots, \mathrm{Rank}(\mathcal{P}^{(d)}))$, where each $\mathcal{P}^{(k)} \in \mathbb{R}^{I_k \times \prod_{m \neq k} I_m}$ is the mode-k expansion of the tensor $\mathcal{P}$ and $\mathrm{Rank}(\mathcal{P}^{(k)})$ denotes the matrix rank of $\mathcal{P}^{(k)}$ [11, 15, 38]. See Supplement for definition of the mode-k expansion. If the Tucker rank of a tensor $\mathcal{P}$ is $(r_1, \ldots, r_d)$, it can be always decomposed as

$$\mathcal{P} = \sum\nolimits_{i_1=1}^{r_1} \cdots \sum\nolimits_{i_d=1}^{r_d} \mathcal{G}_{i_1,\ldots,i_d} \boldsymbol{a}_{i_1}^{(1)} \otimes \boldsymbol{a}_{i_2}^{(2)} \otimes \cdots \otimes \boldsymbol{a}_{i_d}^{(d)}$$

with a tensor $\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_d}$, called the core tensor of $\mathcal{P}$, and vectors $\boldsymbol{a}_{i_k}^{(k)} \in \mathbb{R}^{I_k}$, $i_k \in \{1, \ldots, r_k\}$, for each $k \in \{1, \ldots, d\}$, where $\otimes$ denotes the Kronecker product [24].

If every element of the Tucker rank is the same with each other, it coincides with the CP rank. In this paper, we say that a tensor is rank-1 if its Tucker rank is $(1, \ldots, 1)$. The problem of non-negative Tucker rank reduction is approximating a given non-negative tensor by a non-negative lower Tucker rank tensor. In this paper, Ω^* means the complementary set of a set Ω. We denote by $[d] = \{1, 2, \ldots, d\}$ for a positive integer d and denote by $\mathcal{P}_{a^{(k)}:b^{(k)}}$ the subtensor obtained by fixing the range of kth index to only from a to b.

2.2 The LTR Algorithm

In LTR, we use the rank-1 approximation method that always finds the rank-1 tensor that minimizes the KL divergence from an input tensor [19]. The optimal rank-1 tensor $\overline{\mathcal{P}}$ of $\mathcal{P}$ is given by

$$\overline{\mathcal{P}} = \lambda \boldsymbol{s}^{(1)} \otimes \boldsymbol{s}^{(2)} \otimes \cdots \otimes \boldsymbol{s}^{(d)}, \tag{1}$$

where each $\boldsymbol{s}^{(k)} = (s_1^{(k)}, \dots, s_{I_k}^{(k)})$ with $k \in [d]$ is defined as

$$s_{i_k}^{(k)} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_{k-1}=1}^{I_{k-1}} \sum_{i_{k+1}=1}^{I_{k+1}} \cdots \sum_{i_d=1}^{I_d} \mathcal{P}_{i_1,\dots,i_d}$$

and λ is the $(1-d)$-th power of the total sum of $\mathcal{P}$.

Now we introduce LTR, which iteratively applies the above rank-1 approximation to subtensors of a tensor $\mathcal{P} \in \mathbb{R}^{I_1 \times \cdots \times I_d}$. When we reduce the Tucker rank of $\mathcal{P}$ to $(r_1, \dots, r_d)$, LTR performs the following two steps for each $k \in [d]$:

Step 1: We construct $C = \{c_1, \dots, c_{r_k}\} \subseteq [I_k]$ by random sampling from $[I_k]$ without replacement, where we always assume that $c_1 = 1$ and $c_l < c_{l+1}$ for every $l \in [r_k - 1]$.

Step 2: For each $l \in [r_k]$, if $c_l \neq c_{l+1} - 1$ holds, we replace the subtensor $\mathcal{P}_{c_l^{(k)}:c_{l+1}^{(k)}-1}$ of $\mathcal{P}$ by its rank-1 approximation obtained by Equation (1).

Since the above procedure repeats the best rank-1 approximation $r_1 r_2 \dots r_d$ times at most, the worst computational complexity of LTR is $O(r_1 r_2 \dots r_d I_1 I_2 \dots I_d)$. The choice of C in **Step 1** is arbitrary, thus one can use another strategy, for example, choosing uninformative indices according to prior knowledge if availabe.

3 Theoretical Analysis of LTR

To theoretically guarantee that LTR always reduces the Tucker rank to $(r_1, \dots, r_d)$, in the following subsections, we introduce information geometric analysis of the low-rank tensor approximation by treating each tensor as a probability distribution. The proof of the propositions are given in the supplementary material.

3.1 Modeling tensors as probability distributions

Our key idea for the derivation of LTR is the information geometric treatment of positive tensors in a probability space. We use the special case of a log-linear model on partially ordered sets (poset) introduced by Sugiyama et al. [34] for our probabilistic formulation as they have studied tensors on a statistical manifold [35] using the formulation. We see that the low Tucker rank tensor approximation for positive tensors in terms of the KL divergence is formulated as a convex optimization problem.

In the following, we assume that an input positive tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1 \times \cdots \times I_d}$ is always normalized so that the sum of all elements is 1, and we regard $\mathcal{P}$ as a discrete probability distribution whose sample space is the index set of tensors $\Omega_d = [I_1] \times \cdots \times [I_d]$. Any positive normalized tensor can be described by canonical parameters $(\theta)_{i_1,\dots,i_d} = (\theta_{1,\dots,1}, \dots, \theta_{I_1,\dots,I_d})$ as

$$\mathcal{P}_{i_1,\dots,i_d} = \exp\left(\sum\nolimits_{i'_1=1}^{i_1} \cdots \sum\nolimits_{i'_d=1}^{i_d} \theta_{i'_1,\dots,i'_d} \right). \tag{2}$$

The condition of normalization is exposed on $\theta_{1,\dots,1}$ with $\Omega_d^+ = \Omega_d \backslash \{(1,\dots,1)\}$ as

$$\theta_{1,\dots,1} = -\log \left(\sum_{(i_1,\dots,i_d) \in \Omega_d^+} \exp\left(\sum\nolimits_{i'_1=1}^{i_1} \cdots \sum\nolimits_{i'_d=1}^{i_d} \theta_{i'_1,\dots,i'_d} \right) \right). \tag{3}$$

This belongs to the log-linear model and a parameter vector $(\theta)_{i_1,\dots,i_d}$ in Equation (2) uniquely identifies the normalized positive tensor $\mathcal{P}$. Therefore $(\theta)_{i_1,\dots,i_d}$ can be used as alternative representation of $\mathcal{P}$. If we see each probabilistic distribution, a normalized tensor in our case, as a point in a statistical manifold, $(\theta)_{i_1,\dots,i_d}$ corresponds to a coordinate of the manifold, which is a typical approach in information geometry and known as the *θ-coordinate system* [2].

It is known that any distribution in an exponential family $\log p_\theta(x) = C(x) + \sum_{i=1}^k \theta_i F_i(x) - \psi(\theta)$ for the normalization factor $\psi(\theta)$ and canonical parameters $(\theta)_i$ can be also uniquely identified by expectation parameters $\eta_i = \mathbb{E}_{p_\theta}[F_i]$ [4]. In our modeling in Equation (2), which clearly belongs to

the exponential family, each value of the vector of η-parameters $(\eta)_{i_1,\ldots,i_d}$ is written as follows and it uniquely identifies a normalized positive tensor $\mathcal{P}$:

$$\eta_{i_1,\ldots,i_d} = \sum\nolimits_{i'_1=i_1}^{I_1} \cdots \sum\nolimits_{i'_d=i_d}^{I_d} \mathcal{P}_{i'_1,\ldots,i'_d}. \tag{4}$$

Hence we can also use $(\eta)_{i_1,\ldots,i_d}$ as a coordinate system of the set of distributions, known as the *η-coordinate system* in information geometry [2]. As shown in Sugiyama et al. [34], by using the Möbius function [31, 33] inductively defined as

$$\mu_{i_1,\ldots,i_d}^{i'_1,\ldots,i'_d} = \begin{cases} 1 & \text{if } i_k = i'_k \text{ for all } k \in [d], \\ -\prod_{k=1}^{d} \sum_{j_k=i_k}^{i'_k-1} \mu_{i_1,\ldots,i_d}^{j_1,\ldots j_d} & \text{if } i_k < i'_k \text{ for all } k \in [d], \\ 0 & \text{otherwise,} \end{cases}$$

each distribution $\mathcal{P}$ can be described as

$$\mathcal{P}_{i_1,\ldots,i_d} = \sum\nolimits_{(i'_1,\ldots,i'_d)\in\Omega_d} \mu_{i_1,\ldots,i_d}^{i'_1,\ldots,i'_d} \eta_{i'_1,\ldots,i'_d}. \tag{5}$$

using the η-coordinate system. The normalization condition is realized as $\eta_{1,\ldots,1} = 1$.

In information geometry, it is known that θ- and η-coordinates are connected via Legendre transformation [2]. The remarkable property of the pair of these two coordinate systems is that they are orthogonal with each other and we can combine them to define the (θ, η)-coordinate as a mixture coordinate. If we specify *either* $\theta_{i_1,\ldots,i_d}$ *or* $\eta_{i_1,\ldots,i_d}$ for every $(i_1, \ldots, i_d) \in \Omega_d$, we can always uniquely identify a normalized positive tensor $\mathcal{P}$.

3.2 Bingo rule: θ representation of Tucker rank condition

We theoretically derive the relationship between the Tucker rank of tensors and conditions on the θ-coordinate system. We achieve Tucker rank reduction using such conditions instead of directly imposing constraints on elements of tensors and solving it using classical methods such as the Lagrange multipliers. The advantage of our approach is that we can formulate low-rank approximation as projection of a distribution $\mathcal{P}$ on a subspace of the (θ, η)-space, which always becomes convex optimization due to the flatness of its subspace [2].

Definition 1 (Bingo)**.** *Let $(\theta)_{ij}^{(k)} = (\theta_{11}^{(k)}, \ldots, \theta_{I_k K}^{(k)})$ with $K = \prod_{m\neq k} I_m$ be the θ-coordinate representation of the mode-k expansion of a tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1\times\cdots\times I_d}$. If there exists an integer $i \in [I_k] \setminus \{1\}$ such that $\theta_{ij}^{(k)} = 0$ for all $j \in [K] \setminus \{1\}$, we say that $\mathcal{P}$ has a* bingo on mode-k.

Proposition 1 (Bingo and Tucker rank)**.** *If there are b_k bingos on mode-k, it holds that*

$$\mathrm{Rank}(\mathcal{P}^{(k)}) \leq I_k - b_k.$$

Therefore, for any tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1\times\cdots\times I_d}$ such that it has b_k bingos for each kth index, we can always guarantee that its Tucker rank is at most $(I_1 - b_1, \ldots, I_d - b_d)$.

3.3 Projection theory for LTR

We achieve low Tucker-rank approximation by projection onto a submanifold; that is, our proposed algorithm TLR performs projection in the viewpoint of information geometry. In a statistical manifold parameterized by θ- and η-coordinate systems, two types of geodesics, m- and e-geodesics from a point $\mathcal{P}$ to $\mathcal{Q}$ in a manifold can be defined as

$$\{\, \mathcal{R}_t \mid \mathcal{R}_t = (1-t)\mathcal{P} + t\mathcal{Q} \,\}, \quad \{\, \mathcal{R}_t \mid \log \mathcal{R}_t = (1-t)\log \mathcal{P} + t \log \mathcal{Q} - \phi(t) \,\},$$

respectively, where $0 \leq t \leq 1$ and $\phi(t)$ is a normalization factor to keep $\mathcal{R}_t$ to be a distribution. In the above definition, we regard $\mathcal{P}$ and $\mathcal{Q}$ as its corresponding coordinates points $(\theta)_{i_1,\ldots,i_d}$ or $(\eta)_{i_1,\ldots,i_d}$. A subspace is called e-flat when any e-geodesic connecting any two points in a subspace is included in the subspace. The vertical descent of an m-geodesic from a point $\mathcal{P}$ to $\mathcal{Q}$ in an e-flat subspace M_e is called m-projection. Similary, e-projection is obtained when we replace all e with m and m with e. The flatness of subspaces guarantees the uniqueness of the projection destination.

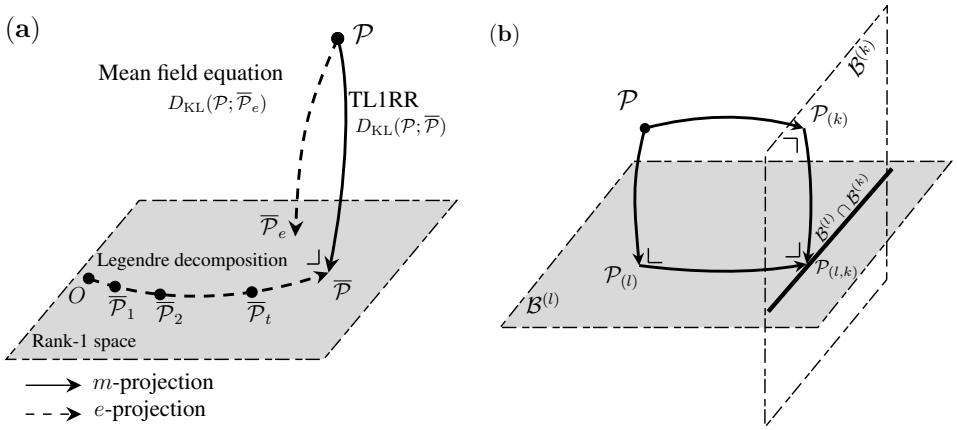

Figure 1: (a) The relationship between rank-1 approximation, Legendre decomposition [35], and mean-field equation, where we assume that the same bingo space is used in Legendre decomposition. A solid line illustrates m-projection with fixing one-body η parameters. $\mathcal{P}$ is an input positive tensor and $\overline{\mathcal{P}}$ is the rank-1 tensor that minimizes the KL divergence from $\mathcal{P}$. $\mathcal{O}$ is an initial point of Legendre decomposition, which is usually a uniform distribution. $\overline{\mathcal{P}}_t$ is a tensor of the t-th step of gradient descent in Legendre Decomposition. (b) The m-projection of a common space of two different bingo spaces from $\mathcal{P}$ can be achieved by m-projection into one bingo space and then m-projecting into the other bingo space.

The projection destination $\mathcal{R}_m$ or $\mathcal{R}_e$ obtained by m- or e-projection onto M_e or M_m minimizes the following KL divergence,

$$\mathcal{R}_m = \operatorname*{argmin}_{\mathcal{Q}\in M_e} D_{\mathrm{KL}}(\mathcal{P};\mathcal{Q}), \quad \mathcal{R}_e = \operatorname*{argmin}_{\mathcal{Q}\in M_m} D_{\mathrm{KL}}(\mathcal{Q};\mathcal{P}).$$

LTR conducts m-projection from an input positive tensor $\mathcal{P}$ onto the low-rank space $\mathcal{B}$, which is the set of tensors each of which satisfies a given set of bingos. In practical situations, a Tucker rank constraint $(r_1, \ldots, r_d)$ is given as an input parameter, and LTR first implicitly constructs $\mathcal{B}$ in **Step 1** by translating the Tucker rank condition into bingos, where any tensor in $\mathcal{B}$ has the Tucker rank at most $(r_1, \ldots, r_d)$. Since the low-rank space $\mathcal{B}$ is e-flat, the m-projection that minimizes the KL divergence is uniquely determined.

In m-projection into a space where some θ-parameters are constrained to be 0, every η-parameter with the same index with some restricted θ-parameter does not change [2]. Therefore, we always have

$$\eta_{i_1,\ldots,i_d} = \eta'_{i_1,\ldots,i_d} \text{ if } \theta'_{i_1,\ldots,i_d} \neq 0, \tag{6}$$

where $\eta_{i_1,\ldots,i_d}$ is the η-coordinate of an input positive tensor $\mathcal{P}$ and $\eta'_{i_1,\ldots,i_d}$ is that of the destination of m-projection onto $\mathcal{B}$ from $\mathcal{P}$. Using the conservation law for the η-coordinate is our key insight to conduct the efficient m-projection onto the bingo space $\mathcal{B}$.

Relationship to Legendre decomposition Legendre decomposition [35] is a tensor decomposition method based on also an information geometric view. However, their concept differs from ours in the following aspect. In the Legendre decomposition, any single point in a subspace that has some constraints on the θ-coordinate is taken as the initial state and moves by gradient descent inside the subspace to minimize the KL divergence from an input tensor. This operation is an e-projection, where the constrained θ-coordinates do not change from the initial state. In contrast, we employ the m-projection from the input tensor onto the low-rank space with fixing some η-coordinates as shown in Equation (6). Using this conservation law for η-coordinates, we can obtain an exact analytical representation of the coordinates of the projection destination without using a gradient method. Figure 1(a) illustrates relationship between our approach and Legendre decomposition. Moreover, the Tucker rank is not discussed in Legendre decomposition, hence it is not guaranteed that Legendre decomposition reduces the Tucker rank, which is in contrast to our method.

3.4 Rank-1 approximation as mean-field approximation

Here we focus on the problem of rank-1 approximation for positive tensors and show the fundamental relationship with the *mean-field theory*. In the following, we refer to the subspace consisting of positive rank-1 tensors as a *rank-1 space*. We use the overline for rank-1 tensors; that is, $\overline{\mathcal{P}}$ is a rank-1 tensor and $\overline{\theta}, \overline{\eta}$ are corresponding parameters of θ- and η-coordinates. Let a *one-body* parameter be a parameter of which at least $d-1$ indices are 1, for example, $\theta_{1,1,3,1}$ and $\eta_{1,5,1,1}$ are one-body parameters for $d=4$. Parameters other than one-body parameters are called *many-body* parameters. These naming come from the Boltzmann machine [1], which is a special case of the log-linear model [34], where a one-body parameter corresponds to a bias and a many-body parameter to a weight. We also use the following notation for one-body parameters of a dth-order tensor,

$$\theta^{(k)}_j \equiv \theta_{\underbrace{1,\dots,1}_{k-1},j,\underbrace{1,\dots,1}_{d-k}}, \quad \eta^{(k)}_j \equiv \eta_{\underbrace{1,\dots,1}_{k-1},j,\underbrace{1,\dots,1}_{d-k}} \quad \text{for each } k \in [d].$$

The rank-1 condition for positive tensors is described as follows using many-body θ parameters as a special case of Proposition 1.

Proposition 2 (rank-1 condition on θ)**.** *For any positive tensor $\overline{\mathcal{P}}$, $\mathrm{rank}(\overline{\mathcal{P}}) = 1$ if and only if its all many-body $\overline{\theta}$ parameters are* 0.

Note that the bingo constraints on θ-coordinates become unique if the target Tucker rank is $(1,\dots,1)$, hence the projection always finds the best rank-1 tensor that minimizes the KL divergence from an input tensor. This means that our information geometric formulation of the rank-1 approximation leads to the same result with the closed formula of the best rank-1 approximation given by Huang and Sidiropoulos [19]. Indeed, using the factorizablity of η-parameters for rank-1 tensors we can reproduce the closed formula in [19] (see Supplement for its proof).

We consider a rank-1 positive tensor $\overline{\mathcal{P}} \in \mathbb{R}^{I_1 \times \cdots \times I_d}_{>0}$ and show that it is represented as a product of independent distributions, which leads to an analogy with the mean-field theory. In the rank-1 space, the normalization condition for dth-order tensors imposed on θ parameters is given as

$$\overline{\theta}_{1,\dots,1} = -\log \prod_{k=1}^{d} \left(1 + \sum_{i_k=2}^{I_k} \exp\left(\sum_{i'_k=2}^{i_k} \overline{\theta}^{(k)}_{i'_k} \right) \right) \tag{7}$$

with assigning 0 to every many-body θ parameter in Equation (3). Note that the empty sum is treated as zero. Next, by substituting 0 for all many-body parameters in our model in Equation (2), we obtain

$$\overline{\mathcal{P}}_{j_1,\dots,j_d} = \exp\left(\overline{\theta}_{1,\dots,1}\right) \prod_{k=1}^{d} \exp\left(\sum_{j'_k=2}^{j_k} \overline{\theta}^{(k)}_{j'_k} \right) \overset{(7)}{=} \prod_{k=1}^{d} \frac{\exp\left(\sum_{j'_k=2}^{j_k} \overline{\theta}^{(k)}_{j'_k}\right)}{1 + \sum_{i_k=2}^{I_k} \exp\left(\sum_{i'_k=2}^{i_k} \overline{\theta}^{(k)}_{i'_k}\right)} \equiv \prod_{k=1}^{d} s^{(k)}_{j_k},$$

where $\boldsymbol{s}^{(k)} \in \mathbb{R}^{I_k}$ is a positive first-order tensor normalized as $\sum_{j_k=1}^{I_k} s^{(k)}_{j_k} = 1$. Since exactly one element is used in each $\boldsymbol{s}^{(k)}$ to determine $\mathcal{P}_{j_1,\dots,j_d}$, we can regard $\boldsymbol{s}^{(k)}$ as a probability distribution with a single random variable $j_k \in [I_k]$. Our above discussion means that any positive rank-1 tensor can be represented as a product of normalized independent distributions.

The operation of approximating a joint distribution as a product of independent distributions is called *mean-field approximation*. The mean-field approximation was invented in physics for discussion of phase transition in ferromagnets [39]. Nowadays, it appears in a wide range of domains such as statistics [30], game theory [6, 28], and information theory [5]. From the viewpoint of information geometry, Tanaka [37] has developed a theory of the mean-field approximation for Boltzmann machines [1], which is defined as $p(\boldsymbol{x}) = \exp(\sum_i b_i x_i + \sum_{ij} w_{ij} x_i x_j)$ for a binary random variable vector $\boldsymbol{x} \in \{0,1\}^n$ with a bias parameter $\boldsymbol{b} = (b)_i \in \mathbb{R}^n$ and an interaction parameter $\boldsymbol{W} = (w_{ij}) \in \mathbb{R}^{n \times n}$. To illustrate that a rank-1 approximation can be regarded as a mean-field approximation, we prepare the mean-field theory of Boltzmann machines in the following.

The mean-field approximation of Boltzmann machines is formulated as the projection from a given distribution onto the e-flat subspace consisting of distributions whose interaction parameters $w_{ij} = 0$ for all i and j, which is called a factorizable subspace. Since the distribution with the constraint

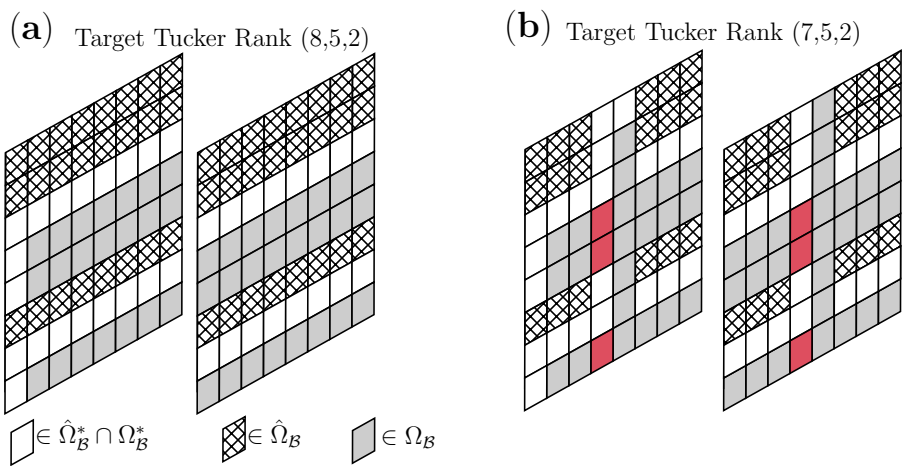

Figure 2: Examples of LTR for $(8, 8, 2)$ tensor. $\Omega_{\mathcal{B}}$ is bingo index. Tensor values and their η-parameters on $\hat{\Omega}_{\mathcal{B}}$ does not change, and η-parameters on $\hat{\Omega}^*_{\mathcal{B}} \cap \Omega^*_{\mathcal{B}}$ also does not change. (a) Target rank is $(8, 5, 2)$. Bingos are on mode-2. In this example, the number of contiguous blocks is 2. (b) Target rank is $(7, 5, 2)$. Bingos are on mode-1 and -2 and the number of contiguous blocks is 1 and 2 for mode-1 and -2, respectively. We assume that we project a tensor onto $\mathcal{B}^{(1)}$, followed by projecting it onto $\mathcal{B}^{(2)}$. After the second m-projection, the θ-parameters on red panels are overwritten. However, these values remain to be zero after the second m-projection in our LTR.

$w_{ij} = 0$ for all i and j can be decomposed into a product of independent distributions, we can approximate a given distribution as a product of independent distribution by the projection onto a factorizable subspace. The m-projection onto the factorizable subspace require to know the expectation value $\eta_i \equiv \mathbb{E}[x_i]$ of an input distributions and it needs $O(2^n)$ computational cost [3], hence we usually approximate it by replacing m-projection with e-projection. The e-projection is usually conducted by a self consistent equation called *mean-field equation*.

The analogy of TL1RR and mean-field theory is clear. In our modeling, a joint distribution $\mathcal{P}$ is approximated by a product of independent distributions $\boldsymbol{s}^{(k)}$ by projecting $\mathcal{P}$ onto the subspace such that all many-body θ parameters are 0, leading to the rank-1 tensor $\overline{\mathcal{P}}$. Since we can compute expectation parameters η by simply summing the input positive tensor in each axial direction, m-projection can be directly performed in our formulation, which is computationally infeasible in the case of Boltzmann machines due to $O(2^n)$ cost. Moreover, the rank-1 space has the same property as factorizable subspace of Boltzmann machines; that is, η can be easily computed from θ (see Supplement for more details). This is the first time that the relationship between mean-field approximation and the tensor rank reduction is demonstrated. The sketch of the relationship among mean-field equations and the rank-1 approximation is illustrated in Figure 1(a).

3.5 Derivation of LTR

Finally, we prove that LTR successfully reduces the Tucker rank by extending the above discussion. We formulate Tucker rank reduction as an m-projection onto a specific bingo space. This bingo space is constructed in **Step 1** of LTR. Then, in **Step 2**, we can archive the m-projection using the closed formula of the rank-1 approximation without a gradient method. We first discuss the case in which the rank of only one mode is reduced, followed by discussing the case in which the rank of two modes are reduced.

When the rank of only one mode is reduced Let us assume that the target Tucker rank is $(I_1, \ldots, I_{k-1}, r_k, I_{k+1}, \ldots, I_d)$ for an input positive tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1 \times \cdots \times I_d}$. Let $\mathcal{B}^{(k)}$ be the set of tensors satisfying bingos and $\Omega_{\mathcal{B}^{(k)}}$ be the set of bingo indices for mode-k constructed in **Step 1** of LTR:

$$\mathcal{B}^{(k)} = \{ \mathcal{P} \mid \theta_{i_1,\ldots,i_d} = 0 \text{ for } (i_1, \ldots, i_d) \in \Omega_{\mathcal{B}^{(k)}} \} . \tag{8}$$

Note that $\mathcal{P} \in \mathcal{B}^{(k)}$ implies that the Tucker rank of $\mathcal{P}$ is at most $(I_1, \ldots, I_{k-1}, r_k, I_{k+1}, \ldots, I_d)$. Let $\tilde{\mathcal{P}}$ be the destination of the m-projection and $\tilde{\theta}$, $\tilde{\eta}$ are its corresponding parameters of θ- and

η-coordinates. From the definition of m-projection and the conservation low of η in Equation (6), the tensor $\tilde{\mathcal{P}}$ should satisfy

$$\tilde{\theta}_{i_1,\dots,i_d} = 0 \text{ for } (i_1,\dots,i_d) \in \Omega_{\mathcal{B}^{(k)}}, \tag{9}$$

$$\tilde{\eta}_{i_1,\dots,i_d} = \eta_{i_1,\dots,i_d} \text{ for } (i_1,\dots,i_d) \notin \Omega_{\mathcal{B}^{(k)}}. \tag{10}$$

As we can see from Equation (5), η-parameters identify a tensor $\mathcal{P}$. In particular, to identify each element $\mathcal{P}_{i_1,\dots,i_d}$, we require η-parameters on only $(i_1,\dots,i_d) \in \{i_1, i_1+1\} \times \{i_2, i_2+1\} \times \cdots \times \{i_d, i_d+1\}$. For example, if $d=2$, $\mathcal{P}_{i_1,i_2} = \eta_{i_1+1,i_2+1} - \eta_{i_1+1,i_2} - \eta_{i_1,i_2+1} + \eta_{i_1,i_2}$ [34]. This leads to the fact that, for $\hat{\Omega}_{\mathcal{B}^{(k)}}$ such that

$$\hat{\Omega}_{\mathcal{B}^{(k)}} = \{\ (i_1,\dots,i_d) \mid \{i_1, i_1+1\} \times \cdots \times \{i_d, i_d+1\} \nsubseteq \Omega_{\mathcal{B}^{(k)}}\ \}, \tag{11}$$

it holds that $\mathcal{P}_{i_1,\dots,i_d} = \tilde{\mathcal{P}}_{i_1,\dots,i_d}$ for $(i_1,\dots,i_d) \in \hat{\Omega}_{\mathcal{B}^{(k)}}$. Therefore, all we have to do to reduce the Tucker rank is to change values for only $(i_1,\dots,i_d) \notin \hat{\Omega}_{\mathcal{B}^{(k)}}$. Such adjustable parts of $\mathcal{P}$ can be divided into some contiguous blocks, and we call each of them a subtensor of $\mathcal{P}$ on mode-k. For example, in Figure 2, we can find two subtensors $\mathcal{P}_{3^{(1)}:5^{(5)}}$ and $\mathcal{P}_{7^{(1)}:8^{(1)}}$. By conducting the rank-1 approximation introduced in Section 2 onto each subtensor, the rank-1 bingo condition ensures the θ condition in Equation (9) and the conservation low of η in Equation (6) for the best rank-1 approximation ensures the η condition in Equation (10). We therefore obtain the tensor $\tilde{\mathcal{P}}$ satisfying Equation (9) and (10) simultaneously by the rank-1 approximations on each subtensor on mode-k, which always belongs to $\mathcal{B}^{(k)}$.

When the rank of only two modes are reduced Let us assume that the target Tucker rank of mode-k is r_k and that of mode-l is r_l. In this case, we need to consider two bingo spaces $\mathcal{B}^{(k)}$ and $\mathcal{B}^{(l)}$ associated with bingo index sets $\Omega_{\mathcal{B}^{(k)}}$ and $\Omega_{\mathcal{B}^{(l)}}$. Let $\mathcal{P}_{(k)}$ be the resulting tensor of m-projection of $\mathcal{P}$ onto $\mathcal{B}^{(k)}$ and $\mathcal{P}_{(k,l)}$ be the resulting tensor of m-projection of $\mathcal{P}_{(k)}$ onto $\mathcal{B}^{(l)}$. Parameters of $\mathcal{P}_{(k,l)}$ satisfy

$$\tilde{\theta}_{i_1,\dots,i_d} = 0 \text{ if } (i_1,\dots,i_d) \in \Omega_{\mathcal{B}^{(k)}} \cup \Omega_{\mathcal{B}^{(l)}}, \tag{12}$$

$$\tilde{\eta}_{i_1,\dots,i_d} = \eta_{i_1,\dots,i_d} \text{ otherwise.} \tag{13}$$

It is obvious that the m-projection from $\mathcal{P}_{(k)}$ onto $\mathcal{B}^{(l)}$ is equivalent to the m-projection from $\mathcal{P}$ onto $\mathcal{B} = \mathcal{B}^{(l)} \cap \mathcal{B}^{(k)}$ The conservation low of η-parameters ensures Equation (12) after the rank-1 approximation of subtensors of $\mathcal{P}_{(k)}$ in the mode-l. In this operation, the part of θ-parameters which are set to be 0 in the previous m-projection onto $\mathcal{B}^{(k)}$ from $\mathcal{P}$ seems to be overwritten (see red panels in Figure 2(b)). However, as shown in Proposition 6 in Supplement, when the rank-1 approximation of a tensor where some one-body θ-parameters are already zero, the values of such one-body θ-parameters after the rank-1 approximation remains to be zero. As a result, Equation (12) holds; that is, we can obtain the tensor $\mathcal{P}_{(k,l)}$ satisfying Equation (12) and (13) simultaneously by the rank-1 approximations on each subtensor of $\mathcal{P}_{(k)}$ on mode-l. We can also immediately confirm $\mathcal{P}_{(l,k)} = \mathcal{P}_{(k,l)}$; that is, the projection order does not matter. The projection sketch is shown in Figure 1(b).

Using the above discussion, we can derive **Step 2** for the general case of Tucker rank reduction.

Theorem 1 (LTR)**.** *For a positive tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1 \times \cdots \times I_d}$, the m-projection destination onto the bingo space $\mathcal{B} = \mathcal{B}^{(1)} \cap \cdots \cap \mathcal{B}^{(d)}$ is given as iterative application of m-projection d times, starting from $\mathcal{P}$ onto subspace $\mathcal{B}^{(1)}$, then from there onto subspace $\mathcal{B}^{(2)}$, …, and finally onto subspace $\mathcal{B}^{(d)}$.*

Since the m-projection minimizes the KL divergence from the input onto the bingo space, LTR always provides the best low-rank approximation in the specified bingo space. The result does not depends on the projection order. The conservation law of η-parameters during the m-projection ensures also that LTR conserves the sum in each axial direction of the input tensor (See Supplement for its proof).

4 Numerical Experiments

We empirically examine the efficiency and the effectiveness of LTR using synthetic and real-world datasets. We compare LTR with two existing non-negative low-tucker rank approximation methods.

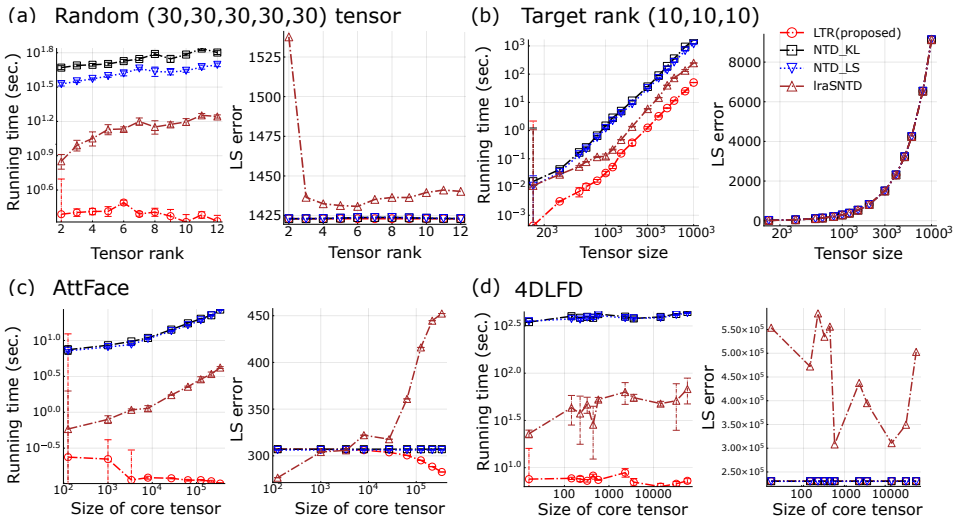

Figure 3: Experimental results for synthetic (a, b) and real-world (c, d) datasets. Means $\pm$ standard error for 20 times iterations are plotted. (a) The horizontal axis is r for target tensor rank (r, r, r, r, r). (b) The horizontal axis is n^3 for input (n, n, n) tensor.

The first one is the non-negative Tucker decomposition, which is the standard nonnegative tensor decomposition method [40] whose cost function is either the Least Squares (LS) error (NTD_LS) or the KL divergence (NTD_KL). The second one is the sequential nonnegative Tucker decomposition (lraSNTD), which is known as a faster approximation method of Tucker decomposition [42]. Its cost function is the LS error. See Supplement for implementation details. We also obtained almost the same results as in Figure 3 with the KL reconstruction error (see Supplement).

Results on Synthetic Data. We create tensors with $d = 3$ or $d = 5$, where every $I_k = n$. We change n to generate various sizes of tensors. Each element is sampled from the uniform continuous distribution on $[0, 1]$. To evaluate the efficiency, we measure the running time of each method. To evaluate the accuracy, we measure the LS reconstruction error. It is defined as the Frobenius norm between input and output tensors. Figure 3(**a**) shows the running time and the LS reconstruction error for randomly generated tensors with $d = 3$ and $n = 30$ with varying the target Tucker tensor rank. Figure 3(**b**) shows running time and the LS reconstruction error for the target tucker rank $(10, 10, 10)$ with varying the input tensor size n. These plots clearly show that our method is faster than other methods with keeping the competitive approximation accuracy.

Results on Real Data. We evaluate running time and the LS reconstruction error for two real world datasets. 4DLFD is a $(9, 9, 512, 512, 3)$ tensor [18] and AttFace is a $(92, 112, 400)$ tensor [32]. AttFace is commonly used in tensor decomposition experiments [21, 22, 42]. See dataset details in Supplement. In both datasets, LTR is always faster than comparison methods as shown in Figure 3(**c, d**) with competitive or better approximation accuracy in terms of the LS error.

5 Conclusion

We have derived a new probabilistic perspective to rank-1 approximation for tensors using information geometry and shown that is can be viewed as mean-field approximation. Our new geometric understanding leads to a novel fast non-negative low Tucker rank approximation method, called LTR, which does not use any gradient method. Our research will not only lead to applications of faster tensor decomposition but also can be a milestone of the research of tensor decomposition to further development of interdisciplinary field around information geometry and the mean-field theory.

This study is theoretical analysis of tensors and we believe that our theoretical discussion would not have negative societal impacts.

References

[1] Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169.

[2] Amari, S. (2016). *Information Geometry and Its Applications*. Springer.

[3] Anderson, J. R. and Peterson, C. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019.

[4] Arwini, K. A. and Dodson, C. T. (2008). *Information geometry*. Springer.

[5] Bhattacharyya, C. and Keerthi, S. S. (2000). Information geometry and Plefka's mean-field theory. *Journal of Physics A: Mathematical and General*, 33(7):1307.

[6] Caines, P. E., Huang, M., and Malhamé, R. (2006). Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 6(3):221–252.

[7] da Silva, A. P., Comon, P., and de Almeida, A. L. (2016). A finite algorithm to compute rank-1 tensor approximations. *IEEE Signal Processing Letters*, 23(7):959–963.

[8] da Silva, A. P., Comon, P., and de Almeida, A. L. F. (2015). Rank-1 tensor approximation methods and application to deflation. *arXiv preprint arXiv:1508.05273*.

[9] De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000a). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278.

[10] De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000b). On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications*, 21(4):1324–1342.

[11] De Silva, V. and Lim, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127.

[12] Dong, W., Li, G., Shi, G., Li, X., and Ma, Y. (2015). Low-rank tensor approximation with laplacian scale mixture modeling for multiframe image denoising. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 442–449.

[13] Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

[14] Grasedyck, L., Kressner, D., and Tobler, C. (2013). A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78.

[15] Hackbusch, W. (2019). *Tensor spaces and numerical tensor calculus*. Number v. 56 in Springer series in computational mathematics. Springer, 2nd ed edition.

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

[17] Hitchcock, F. L. (1928). Multiple invariants and generalized rank of a p-way matrix or tensor. *Journal of Mathematics and Physics*, 7(1-4):39–79.

[18] Honauer, K., Johannsen, O., Kondermann, D., and Goldluecke, B. (2016). A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer.

[19] 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*, pages 693–697. IEEE.

[20] Ji, Y., Wang, Q., Li, X., and Liu, J. (2019). A survey on tensor techniques and applications in machine learning. *IEEE Access*, 7:162950–162990.

[21] Kim, Y.-D. and Choi, S. (2007). Nonnegative tucker decomposition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

[22] 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*, pages 1829–1832. IEEE.

[23] Klus, S. and Gelß, P. (2019). Tensor-based algorithms for image classification. *Algorithms*, 12(11):240.

[24] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

[25] Lathauwer, L. D. (2000). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278.

[26] Lee, D. and Seung, H. S. (2001a). Algorithms for non-negative matrix factorization. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13, pages 556–562. MIT Press.

[27] Lee, D. D. and Seung, H. S. (2001b). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

[28] Lions, P.-L. and Lasry, J.-M. (2007). Large investor trading impacts on volatility. *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*, 24(2):311–323.

[29] Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59.

[30] Peterson, C. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems*, pages 995–1019.

[31] Rota, G.-C. (1964). On the foundations of combinatorial theory I: Theory of Möbius functions. *Z. Wahrseheinlichkeitstheorie*, 2:340–368.

[32] Samaria, F. S. and Harter, A. C. (1994). Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE workshop on applications of computer vision*, pages 138–142. IEEE.

[33] Spiegel, E. and O'Donnell, C. J. (1997). *Incidence Algebras*. Marcel Dekker.

[34] Sugiyama, M., Nakahara, H., and Tsuda, K. (2017). Tensor balancing on statistical manifold. *arXiv preprint arXiv:1702.08142*.

[35] Sugiyama, M., Nakahara, H., and Tsuda, K. (2019). Legendre decomposition for tensors. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124017.

[36] Symeonidis, P., Nanopoulos, A., and Manolopoulos, Y. (2008). Tag recommendations based on tensor dimensionality reduction. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 43–50.

[37] Tanaka, T. (1999). A theory of mean field approximation. In *Advances in Neural Information Processing Systems*, pages 351–360.

[38] Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

[39] Weiss, P. (1907). L'hypothèse du champ moléculaire et la propriété ferromagnétique. *Journal de Physique Théorique et Appliquée*, 6(1):661–690.

[40] Welling, M. and Weber, M. (2001). Positive tensor factorization. *Pattern Recognition Letters*, 22(12):1255–1261.

[41] Zhang, T. and Golub, G. H. (2001). Rank-one approximation to high order tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(2):534–550.

[42] Zhou, G., Cichocki, A., and Xie, S. (2012). Fast nonnegative matrix/tensor factorization based on low-rank approximation. *IEEE Transactions on Signal Processing*, 60(6):2928–2940.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Scope and contributions are clearly described in Section 1.
   (b) Did you describe the limitations of your work? [Yes] The limitations are clearly described in Section 1.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All assumptions are described in Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] We have included them in Supplement.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All the data and code necessary for the reproduction experiment are attached as Supplement.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] Our method is not based on training.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Error bar is plotted in Section 4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We have stated in Supplement.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]