Optimal Submanifold Structure in Log-linear Models

Derun Zhou[1,2] **Mahito Sugiyama**[1,2]

[1]National Institute of Informatics, Tokyo, Japan
[2]The Graduate University for Advanced Studies, SOKENDAI

Abstract

In the modeling of discrete distributions using log-linear models, the model selection process is equivalent to imposing zero-value constraints on a subset of natural parameters, which is an established concept in information geometry. This *zero-value constraint* has been implicitly employed, from classic Boltzmann machines to recent many-body approximations of tensors. However, in theory, any constant value other than zero can be used for these constraints, leading to different *submanifolds* onto which the empirical distribution is projected, a possibility that has not been explored. Here, we investigate the asymptotic behavior of these constraint values from the perspective of information geometry. Specifically, we prove that the optimal value converges to zero as the size of the support of the empirical distribution increases, which corresponds to the size of the input tensors in the context of tensor decomposition. While our primary focus is on many-body approximation of tensors, it can serve as a basis for extending to a wide range of log-linear modeling applications.

1 INTRODUCTION

The energy-based model is widely used in machine learning areas [LeCun et al., 2006, Jaynes, 1957]. Since the exponential family is in the energy-based model, it covers a wide variety of classical distributions for continuous variables, such as Gaussian, exponential, and gamma distributions [MacKay, 2003]. Moreover, the *log-linear model*, which is also in the exponential family, covers all the positive probability distributions over a finite space [Shpitser et al., 2013]. Recently, the log-linear model has been used to model distributions over partially ordered sets (posets) and its dually-flat manifold structure has been analyzed in an information geometric manner [Sugiyama et al., 2017].

The log-linear model on a partial-order structure (LPS) provides an alternative approach to decomposition for positive tensors, which avoids the optimization difficulties associated with the common low-rank based decompositions by replacing the squared error loss with the Kullback-Leibler (KL) divergence [Sugiyama et al., 2018]. Each positive tensor is treated as a discrete distribution with a partial-order structure. It is parameterized by the natural parameters of the exponential family, and the optimization is realized as a projection onto a model submanifold constrained by a subset of these natural parameters. We can capture the nontrivial structure of positive tensors after the projection, one of which is the *many-body tensor approximation* that captures a hierarchy of mode interactions [Ghalamkari et al., 2023]. The mode interaction selection based on many-body approximation can be regarded as the feature selection in distribution learning.

In this paper, we focus on many-body approximation, as it is not only a key application of LPS but also includes a wide variety of graphical models such as standard and high-order Boltzmann machines [Ackley et al., 1985, Sejnowski, 1986]. In many-body approximation, specifying the model submanifold, which can be viewed as a model selection problem or hyperparameter tuning, typically involves imposing the zero-value constraint on a subset of the natural parameters. In Boltzmann machines, this process corresponds to selecting a graphical model, where a zero-value constraint is implicitly applied. Specifically, removing an edge between nodes, each of which represents a random variable, effectively sets the corresponding natural parameter to zero. However, from an information geometric perspective, these constraints could, in principle, take any constant value other than zero. Despite this flexibility, this possibility remains largely unexplored.

We provide a formal description and a simple example below. In the modeling based on the LPS, including many-body approximation and Boltzmann machines, we first select the model submanifold, also known as an e-flat submanifold, described as $\mathcal{S}_{\mathcal{B}}^{0} = \left\{\mathcal{Q} \in \mathcal{S} \mid \theta_v = 0 \text{ for all } v \in \Omega_d^{+} \backslash \mathcal{B}\right\}$, where $\mathcal{S}$ denotes the set of distributions, and θ_v represents the natural parameters of the LPS (exponential family). We define $\Omega_d = [I_1] \times \cdots \times [I_d]$, where $[I_k] = \{1, 2, \ldots, I_k\}$. To exclude the normalization constant, we often work in the reduced space $\Omega_d^{+} = \Omega_d \setminus \{(1, 1, \ldots, 1)\}$, and consider a subset $\mathcal{B} \subseteq \Omega_d^{+}$. The parameters in $\mathcal{B}$ are optimized by minimizing the KL divergence.

Here it is clear from the equation that this model submanifold allows not only $\mathcal{S}_{\mathcal{B}}^{0}$ with the "$\theta_v = 0$" constraint but also $\mathcal{S}_{\mathcal{B}}^{c}$ with the "$\theta_v = c$" constraint for any constant value c, which may help to decrease the KL error further. For example, let us consider decomposing the following toy matrix:

$$\begin{bmatrix} 833 & 1 & 2 & 4 & 7 & 4 & 8 \\ 430 & 33 & 5 & 1 & 711 & 112 & 4 \\ 39 & 6 & 29 & 2 & 9 & 3 & 121 \\ 2 & 2 & 8 & 6 & 311 & 10 & 122 \end{bmatrix}.$$

We choose the decomposition basis as one body natural parameters, which means θ_{1j} and θ_{i1} are selected as decomposition basis. If we choose the submanifold as $\mathcal{S}_{\mathcal{B}_1}^{0}$, where $\mathcal{B}_1$ is the index set of one-body natural parameters, the KL error is 0.46 and the RMSE is 0.56, and the projection result is a rank-1 matrix as follows:

$$\begin{bmatrix} 396.5 \\ 598.2 \\ 96.5 \\ 212.8 \end{bmatrix} \begin{bmatrix} 1.0 & 0.03 & 0.03 & 0.01 & 0.8 & 0.1 & 0.2 \end{bmatrix}$$
$$= \begin{bmatrix} 396.5 & 11.9 & 11.9 & 4.0 & 317.2 & 39.7 & 79.3 \\ 598.2 & 17.95 & 17.95 & 5.98 & 478.6 & 59.8 & 119.6 \\ 96.5 & 2.9 & 2.9 & 0.97 & 77.2 & 9.65 & 19.3 \\ 212.8 & 6.38 & 6.38 & 2.13 & 170.2 & 21.3 & 42.56 \end{bmatrix}.$$

In contrast, if we choose $\mathcal{S}_{\mathcal{B}_1}^{0.54}$ as the model submanifold, the resulting KL error is 0.19 and RMSE is only 0.24, which is a half of the result of $\mathcal{S}_{\mathcal{B}_1}^{0}$. The reconstruction matrix is in the following.

$$\begin{bmatrix} 731.1 & 17.6 & 12.5 & 2.2 & 88.7 & 4.3 & 2.55 \\ 555.05 & 22.96 & 27.99 & 8.49 & 583.33 & 48.7 & 49.49 \\ 14.95 & 1.06 & 2.22 & 1.16 & 136.1 & 19.5 & 34 \\ 2.9 & 0.35 & 1.27 & 1.14 & 229.9 & 56.5 & 169 \end{bmatrix}.$$

Please note that it is no longer rank-1, while the number of free parameters is the same with the case of $\mathcal{S}_{\mathcal{B}_1}^{0}$. This example highlights the necessity of studying the submanifold selection problem. For the detail explanation of this example, please refer to Appendix A.5.

To summarize, our contribution is threefold:

- We theoretically prove that, for any order many-body approximation, the optimal e-flat model submanifold converges to $\mathcal{S}_{\mathcal{B}}^{0}$ as the tensor size (the number of entries of a tensor) increases.
- We present an optimal e-flat submanifold searching algorithm. This algorithm is formulated as a convex optimization, hence it always finds the globally optimal solution of a KL divergence minimization problem with linear constraint conditions. This algorithm can be used to improve the performance of small or medium-scale datasets for tensor decomposition or distribution learning for tabular data.
- We provide and empirical evaluation on synthetic and real-world datasets and show the consistency between theory and experimental results.

2 PRELIMINARIES

2.1 FORMULATION

We start with a positive d^{th}-order input tensor $\mathcal{X} \in \mathbb{R}_{>0}^{I_1 \times \cdots \times I_d}$ and normalize it as $\hat{\mathcal{P}}_{i_1,\ldots,i_d} = \mathcal{X}_{i_1,\ldots,i_d} / \sum_{j_1=1}^{I_1} \cdots \sum_{j_d=1}^{I_d} \mathcal{X}_{j_1,\ldots,j_d}$. For the remainder of this paper, we consistently work on the normalized tensor $\hat{\mathcal{P}}$ as the input tensor. We can treat any normalized tensor as a discrete distribution (or a probabilistic vector) with the sample space $\Omega_d = [I_1] \times \cdots \times [I_d]$, where $[I_k] = \{1, 2, \ldots, I_k\}$. Hence, it is exactly modeled by the *log-linear model*:

$$\log \mathcal{P}_{i_1,\ldots,i_d} = \sum_{i_1'=1}^{i_1} \cdots \sum_{i_d'=1}^{i_d} \theta_{i_1',\ldots,i_d'} \tag{1}$$

for each $(i_1, \ldots, i_d) \in \Omega_d$, where each $\theta_{i_1',\ldots,i_d'} \in \mathbb{R}$ corresponds to a *natural parameter*. The normalization is exposed on $\theta_\perp$ with $\perp = (1, \ldots, 1)$ as

$$\theta_\perp = -\log \left(\sum_{(i_1,\ldots,i_d) \in \Omega_d^{+}} \exp \left(\sum_{i_1'=1}^{i_1} \cdots \sum_{i_d'=1}^{i_d} \theta_{i_1',\ldots,i_d'} \right) \right). \tag{2}$$

For example of *log-linear model*, please refer to Appendix A.6. Thus we often work on the space $\Omega_d^{+} = \Omega_d \backslash \{(1, 1, \ldots, 1)\}$ by excluding the normalization constant. In addition to natural parameters, we also have another set of parameters called expectation parameters, denoted as a vector $(\eta)_{i_1,\ldots,i_d}$. Each value of the η-parameter vector is written as follows:

$$\eta_{i_1,\ldots,i_d} = \sum_{i_1'=i_1}^{I_1} \cdots \sum_{i_d'=i_d}^{I_d} \mathcal{P}_{i_1',\ldots,i_d'}, \tag{3}$$

and uniquely identifies a normalized positive tensor $\mathcal{P}$ by the following equation.

$$\mathcal{P}_{i_1,\ldots,i_d} = \sum_{(i_1',\ldots,i_d') \in \Omega_d} \mu_{i_1,\ldots,i_d}^{i_1',\ldots,i_d'} \eta_{i_1',\ldots,i_d'}, \tag{4}$$

where μ is the Möbius function defined inductively as

$$\mu_{i_1,\dots,i_d}^{i'_1,\dots,i'_d} = \begin{cases} 1 & i_k = i'_k, \forall k \in [d], \\ -\prod_{k=1}^{d} \sum_{j_k=i_k}^{i'_k-1} \mu_{i_1,\dots,i_d}^{j_1,\dots j_d} & i_k < i'_k, \forall k \in [d], \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

An example of Equation (4) is presented in Appendix A.4. The normalization condition is realized as $\eta_{1,\dots,1} = 1$. Both of $(\theta)_{i_1,\dots,i_d}$ and $(\eta)_{i_1,\dots,i_d}$ serve as coordinate systems for the set of distributions.

2.2 LEGENDRE DECOMPOSITION

We introduce the Legendre decomposition [Sugiyama et al., 2018], which decomposes a given tensor via log-linear modeling introduced in the previous subsection. Let $\mathcal{S}$ be the set of all normalized positive tensors. When we have an index set $\mathcal{B} \subseteq \Omega_d^+$ as a *decomposition basis*, the corresponding submanifold $\mathcal{S}_{\mathcal{B}}^0$ is given as

$$\mathcal{S}_{\mathcal{B}}^0 = \left\{ Q \in \mathcal{S} \mid \theta_{i_1,\dots,i_d} = 0 \text{ for all } (i_1,\dots,i_d) \in \Omega_d^+ \backslash \mathcal{B} \right\}.$$

Legendre decomposition is formulated as optimization that finds $\mathcal{P}^{\mathcal{B},0}$ in the submanifold $\mathcal{S}_{\mathcal{B}}^0$ minimizing the following KL divergence:

$$\mathcal{P}^{\mathcal{B},0} = \operatorname*{argmin}_{\mathcal{R} \in \mathcal{S}_{\mathcal{B}}^0} D_{\mathrm{KL}}(\hat{\mathcal{P}}, \mathcal{R}),$$

where the KL divergence from $\hat{\mathcal{P}} \in \mathcal{S}$ to $\mathcal{R} \in \mathcal{S}$ is given as

$$D_{KL}(\hat{\mathcal{P}}, \mathcal{R}) = \sum_{i_1=1}^{I_1} \cdots \sum_{i_d=1}^{I_d} \hat{\mathcal{P}}_{i_1,\dots,i_d} \log \frac{\hat{\mathcal{P}}_{i_1,\dots,i_d}}{\mathcal{R}_{i_1,\dots,i_d}}.$$

It is known that the derivative of the KL divergence is

$$\frac{\partial}{\partial \theta_{i_1,\dots,i_d}} D_{KL}(\hat{\mathcal{P}}, \mathcal{R}) = \eta_{i_1,\dots,i_d} - \hat{\eta}_{i_1,\dots,i_d} \tag{6}$$

for every $(i_1,\dots,i_d) \in \mathcal{B}$, where $(\eta)_{i_1,\dots,i_d}$ and $(\hat{\eta})_{i_1,\dots,i_d}$ are the expectation parameters of $\mathcal{R}$ and $\hat{\mathcal{P}}$, respectively. This equation implies that the KL divergence is minimized if and only if $\eta_{i_1,\dots,i_d} = \hat{\eta}_{i_1,\dots,i_d}$ for all $(i_1,\dots,i_d) \in \mathcal{B}$. In information geometry, this optimization problem can be regarded as the m-projection onto the e-flat submanifold $\mathcal{S}_{\mathcal{B}}^0$. The tensor $\mathcal{P}^{\mathcal{B},0}$, such that $\mathcal{P}^{\mathcal{B},0} \in \mathcal{S}_{\mathcal{B}}^0 \cap \mathcal{S}_{\hat{\mathcal{P}}}^{\mathcal{B}}$, always uniquely exists, where

$$\mathcal{S}_{\hat{\mathcal{P}}}^{\mathcal{B}} = \{ Q \in \mathcal{S} \mid \eta_{i_1,\dots,i_d} = \hat{\eta}_{i_1,\dots,i_d} \text{ for all } (i_1,\dots,i_d) \in \mathcal{B} \}. \tag{7}$$

Moreover, $\mathcal{S}_{\hat{\mathcal{P}}}^{\mathcal{B}}$ is an m-flat submanifold since it imposes constraints on the η coordinate. For the definitions of m-flat and e-flat submanifolds, as well as the concept of projection theory, please refer to Appendix A.1 and A.2.

2.3 MANY-BODY APPROXIMATION

Many-body approximation is a special case of Legendre decomposition, which emphasizes the connection to the mode interactions of tensors by explicitly incorporating them in the modeling [Ghalamkari et al., 2023]. For each $\theta_{i_1,\dots,i_d}$, if there are h non-one indices, we call it an h-body parameter. For example, if we consider a 4th-order input tensor, $\theta_{1,2,1,1}$ is a one-body parameter, $\theta_{4,3,1,1}$ is a two-body parameter, $\theta_{1,2,4,3}$ is a three-body parameter and $\theta_{5,2,4,3}$ is a four-body parameter.

The definition of many-body approximation can be summarized in the following: For a given tensor $\hat{\mathcal{P}}$, its h-body approximation is the optimal solution $\mathcal{P}^{\mathcal{B}_h,0}$ such that

$$\mathcal{P}^{\mathcal{B}_h,0} = \operatorname*{argmin}_{\mathcal{R} \in \mathcal{S}_{\mathcal{B}_h}^0} D_{KL}(\hat{\mathcal{P}}, \mathcal{R}),$$

where the solution space $\mathcal{S}_{\mathcal{B}_h}^0$ is given as $\mathcal{S}_{\mathcal{B}_h}^0 = \{\mathcal{Q} \in \mathcal{S} \mid \theta_{i_1,\dots,i_d} = 0 \text{ if } \theta_{i_1,\dots,i_d} \text{ is } n\ (n > h)\text{-body parameters of } \mathcal{Q}\}$. Therefore, the decomposition basis $\mathcal{B}_h$ is the index set composed of all i-body parameters with $1 \le i \le h$, and the inclusion relationship $\mathcal{B}_h \subseteq \mathcal{B}_{h+1}$ always holds. Moreover, it is important to note that $\mathcal{P}^{\mathcal{B}_d,0} = \hat{\mathcal{P}}$.

3 THEORETICAL ANALYSIS

We theoretically analyze the behavior of $c \in \mathbb{R}$ for the e-flat model submanifold:

$$\mathcal{S}_{\mathcal{B}_h}^c = \left\{ Q \in \mathcal{S} \mid \theta_v = c \text{ for all } v \in \Omega_d^+ \setminus \mathcal{B}_h \right\}$$

in h-body approximation of an input tensor $\hat{\mathcal{P}} \in \mathbb{R}_{>0}^{I_1 \times \cdots \times I_d}$. Compared with $\mathcal{S}_{\mathcal{B}_h}^0$ that we mentioned in the previous section, here c is a constant that is not limited to 0. The result of m-projection of $\hat{\mathcal{P}}$ onto the submanifold $\mathcal{S}_{\mathcal{B}_h}^c$ is still formulated as $\mathcal{P}^{\mathcal{B}_h,c} = \operatorname{argmin}_{\mathcal{R} \in \mathcal{S}_{\mathcal{B}_h}^c} D_{KL}(\hat{\mathcal{P}}, \mathcal{R})$ and, according to the projection theory, it is always guaranteed that $\mathcal{P}^{\mathcal{B}_h,c}$ not only exists but is also unique. For further details, refer to Appendix A.3. The objective of our theoretical analysis is to find out whether there exists an e-flat submanifold $\mathcal{S}_{\mathcal{B}_h}^{c_0}$ and its m-projection result $\mathcal{P}^{\mathcal{B}_h,c_0}$ satisfying

$$D_{\mathrm{KL}}(\hat{\mathcal{P}}, \mathcal{P}^{\mathcal{B}_h,c_0}) \le D_{\mathrm{KL}}(\hat{\mathcal{P}}, \mathcal{P}^{\mathcal{B}_h,c}) \quad \text{for all } c \in \mathbb{R}.$$

This means that there exists an optimal low-dimensional submanifold $\mathcal{S}_{\mathcal{B}_h}^{c_0}$ which ensures that the KL divergence reaches its minimum value under the same dimensionality, under the fixed number h of bodies. As we show in Figure 1, each e-flat model submanifold, $\mathcal{S}_{\mathcal{B}_h}^c$, $\mathcal{S}_{\mathcal{B}_h}^0$, $\mathcal{S}_{\mathcal{B}_h}^{-c}$, and $\mathcal{S}_{\mathcal{B}_h}^{c_0}$, has a unique intersection with the m-flat (data) submanifold $\mathcal{S}_{\hat{\mathcal{P}}}^{\mathcal{B}_h}$, which corresponds to $\mathcal{P}^{\mathcal{B}_h,c}$, $\mathcal{P}^{\mathcal{B}_h,0}$, $\mathcal{P}^{\mathcal{B}_h,-c}$, and $\mathcal{P}^{\mathcal{B}_h,c_0}$, respectively. Please note that $\mathcal{S}_{\hat{\mathcal{P}}}^{\mathcal{B}_h}$ is defined by replacing $\mathcal{B}$ with $\mathcal{B}_h$ in Equation (7).

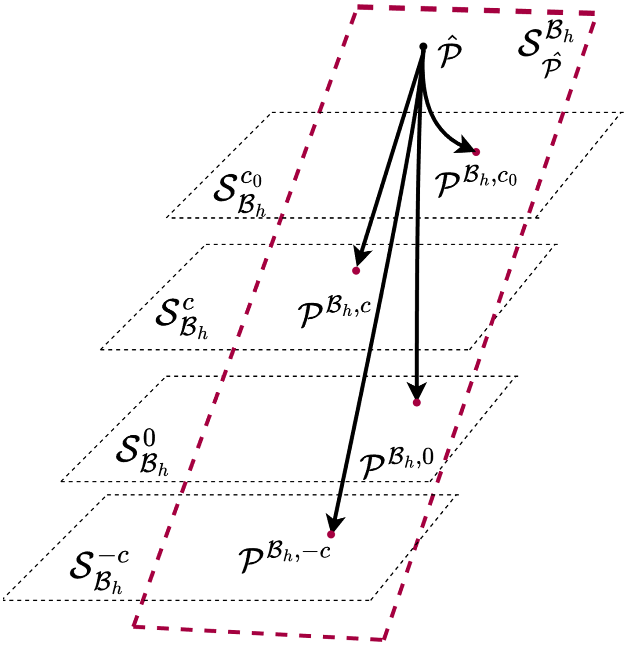

$\hat{\mathcal{P}}$
$\mathcal{S}^{\mathcal{B}_h}_{\hat{\mathcal{P}}}$
$\mathcal{S}^{c_0}_{\mathcal{B}_h}$
$\mathcal{P}^{\mathcal{B}_h,c_0}$
$\mathcal{S}^{c}_{\mathcal{B}_h}$
$\mathcal{P}^{\mathcal{B}_h,c}$
$\mathcal{S}^{0}_{\mathcal{B}_h}$
$\mathcal{P}^{\mathcal{B}_h,0}$
$\mathcal{S}^{-c}_{\mathcal{B}_h}$
$\mathcal{P}^{\mathcal{B}_h,-c}$

Figure 1: The m-projection onto different e-flat model submanifolds.

In the following, we theoretically prove that c_0 exists and converges to 0 as the size of a input tensor $\hat{\mathcal{P}}$ increases. Here, the size of the tensor is defined as the total number of elements in the normalized tensor $\hat{\mathcal{P}}$, simply given by $\prod_{j=1}^{d} I_j$. Specifically, we primarily consider two approaches to increasing the size of the tensor. The first approach involves increasing the values of I_j for each $j = 1, \ldots, d$, while the second approach increases the dimensionality d of the tensor. First, to prove the main result, we derive the closed form of $\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}$, which is the result of m-projection of $\hat{\mathcal{P}}$ onto a special submanifold

$$\mathcal{S}^{c}_{\mathcal{B}_{I_j,m_j}} = \left\{ Q \in \mathcal{S} \mid \theta_v = c \text{ for all } v \in \Omega_d^{+} \backslash \mathcal{B}_{I_j,m_j} \right\}.$$

where $\mathcal{B}_{I_j,m_j} = [I_1]\times\cdots\times[I_{j-1}]\times[m_j]\times[I_{j+1}]\times\cdots\times[I_d]$ and $[m_j] = \{1, \ldots, m_j\}$ with $m_j \leq I_j$. Figure 2 shows an example of the submanifold $\mathcal{S}^{c}_{\mathcal{B}_{I_3,m_3}}$ for a 3^{th}-order tensor.

To obtain the closed formula of $\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}$, first we show the following lemma.

Lemma 1. *For any input tensor $\hat{\mathcal{P}} \in \mathbb{R}_{>0}^{I_1\times\cdots\times I_d}$ and its m-projection $\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}$ onto the submanifold $\mathcal{S}^{c}_{\mathcal{B}_{I_j,m_j}}$, we have*

$$\sum_{i_j=m_j}^{I_j} \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d} = \sum_{i_j=m_j}^{I_j} \hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d}. \tag{8}$$

The proof of Lemma 1 can be found in Appendix B. This lemma indicates that the m-projection onto the submanifold $\mathcal{S}^{c}_{\mathcal{B}_{I_j,m_j}}$ preserves the marginal sum over the j-th mode. Based on the above lemma, we give the following theorem.

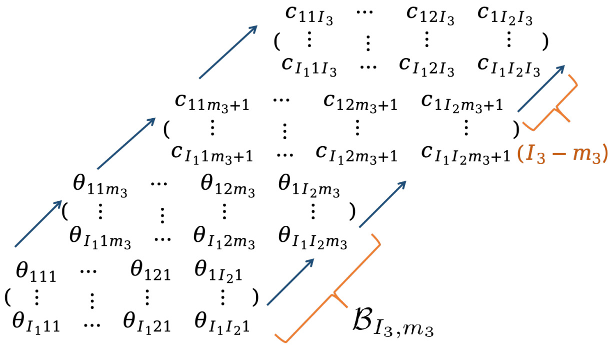

c_{11I_3} ⋯ c_{12I_3} $c_{1I_2I_3}$
$c_{I_11I_3}$ ⋯ $c_{I_12I_3}$ $c_{I_1I_2I_3}$
c_{11m_3+1} ⋯ c_{12m_3+1} $c_{1I_2m_3+1}$
$c_{I_11m_3+1}$ ⋯ $c_{I_12m_3+1}$ $c_{I_1I_2m_3+1}$
$(I_3 - m_3)$
θ_{11m_3} ⋯ θ_{12m_3} $\theta_{1I_2m_3}$
$\theta_{I_11m_3}$ ⋯ $\theta_{I_12m_3}$ $\theta_{I_1I_2m_3}$
θ_{111} ⋯ θ_{121} θ_{1I_21}
θ_{I_111} ⋯ θ_{I_121} $\theta_{I_1I_21}$
$\mathcal{B}_{I_3,m_3}$

Figure 2: The submanifold $\mathcal{S}^{c}_{\mathcal{B}_{I_3,m_3}}$, with $\mathcal{B}_{I_3,m_3} = [I_1] \times [I_2] \times [m_3]$, $[m_3] = \{1, \ldots, m_3\}$.

Theorem 1. *For any input tensor $\hat{\mathcal{P}} \in \mathbb{R}_{>0}^{I_1\times\cdots\times I_d}$, its m-projection $\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}$ onto the submanifold $\mathcal{S}^{c}_{\mathcal{B}_{I_j,m_j}}$ is given as*

$$\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d} = \hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d}$$

for $i_j = 1, 2, \ldots, m_j - 1$ and

$$\begin{aligned}&\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d} \\ &= \left(\sum_{k=0}^{I_j-m_j} \hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d} \right) \\ &\quad \cdot \exp\left(ck \prod_{\substack{s=1\\ s\neq j}}^{d} i_s \right) \left(\sum_{k=0}^{I_j-m_j} \exp\left(ck \prod_{\substack{s=1\\ s\neq j}}^{d} i_s \right) \right)^{-1}\end{aligned} \tag{9}$$

for $k = 0, \ldots, I_j - m_j$.

The proof can be found in Appendix B. Theorem 1 provides the closed formula of $\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}$, hence the closed formulae of $D_{\mathrm{KL}}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c})$ and $D_{\mathrm{KL}}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},0})$ can also be obtained directly, facilitating the estimation of their bounds. Because

$$D_{KL}(\hat{\mathcal{P}}, \mathcal{R}) = \sum_{i_1=1}^{I_1} \cdots \sum_{i_d=1}^{I_d} \hat{\mathcal{P}}_{i_1,\ldots,i_d} \log \frac{\hat{\mathcal{P}}_{i_1,\ldots,i_d}}{\mathcal{R}_{i_1,\ldots,i_d}}$$

and the term $\sum_{i_1=1}^{I_1} \cdots \sum_{i_d=1}^{I_d} \hat{\mathcal{P}}_{i_1,\ldots,i_d} \log \hat{\mathcal{P}}_{i_1,\ldots,i_d}$ is not related to $\mathcal{R}$, we only need to consider

$$F(\hat{\mathcal{P}}; \mathcal{R}) = -\sum_{i_1=1}^{I_1} \cdots \sum_{i_d=1}^{I_d} \hat{\mathcal{P}}_{i_1\ldots i_d} \log \mathcal{R}_{i_1,\ldots,i_d}.$$

Let $s_{\min}$ be the value defined by the equation $\min(\hat{\mathcal{P}}) = 1/(\prod_{j=1}^{d} I_j)^{s_{\min}}$, where the term $\prod_{j=1}^{d} I_j$ is the total number of elements in the normalized tensor $\hat{\mathcal{P}}$. Then it is trivial that $s_{\min} \geq 1$ always holds. Similarly, let $s_{\max}$ be the value satisfying $\max(\hat{\mathcal{P}}) = 1/(\prod_{j=1}^{d} I_j)^{s_{\max}}$. Then we always have $0 < s_{\max} \leq 1$. Furthermore, to facilitate the discussion, we assume that $0 < a \leq \mathcal{X}_{i_1,\dots,i_d} \leq b$ always holds for some constant values a and b and remains unchanged as the tensor size increases. Consequently, we have $\min(\hat{\mathcal{P}}) \geq a/(b\prod_{j=1}^{d} I_j)$ and $\max(\hat{\mathcal{P}}) \leq b/(a\prod_{j=1}^{d} I_j)$.

Here we present the following theorem using the above properties.

Theorem 2. $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq \max\{l, -l\}$, *where*

$$l = \frac{c\,(I_j - m_j)\,(I_j - m_j + 1)\prod_{\substack{h=1\\h\neq j}}^{d} I_h\,(1 + I_h)}{2^d \left(\prod_{j=1}^{d} I_j\right)^{s_{min}}}.$$

The proof of Theorem 2 can be found in Appendix B. In the following, we use $\underline{F}(\cdot)$ and $\overline{F}(\cdot)$ to denote the lower and upper bounds of the function F, respectively. Consequently, this theorem establishes the lower bound of $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c})$, denoted as $\underline{F}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c})$, which enables us to determine the range of c that satisfies $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq \underline{F}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq \overline{F}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},0}) \geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},0})$. Moreover, the range of c satisfying $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \leq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},0})$ is a subset of the complement of the range of c that satisfies $\underline{F}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq \overline{F}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},0})$. Based on the above analysis, the following corollary holds.

Corollary 3. *To satisfy the condition* $D_{\mathrm{KL}}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \leq D_{\mathrm{KL}}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},0})$, *for every* $j = 1, \dots, d$, c *should at least satisfy* $-l \leq c \leq l$, *where*

$$\begin{aligned} l &= \frac{2^d s_{\max} \log\left(\prod_{j=1}^{d} I_j\right)}{(I_j - m_j)\prod_{\substack{h=1\\h\neq j}}^{d} (1 + I_h)\left(\prod_{j=1}^{d} I_j\right)^{s_{\max}-s_{\min}}} \\ &< \frac{2^d \log\left(\prod_{j=1}^{d} I_j\right) b^2}{(I_j - m_j)\prod_{\substack{h=1\\h\neq j}}^{d} (1 + I_h)\, a^2}. \end{aligned} \tag{10}$$

The proof of Corollary 3 can be found in Appendix B. This corollary shows that $\mathcal{P}^{\mathcal{B}_{I_j,m_j},c_0}$ eventually converges to $\mathcal{P}^{\mathcal{B}_{I_j,m_j},0}$ as the tensor size increases. We set $m_j = \lfloor I_j/\alpha \rfloor \geq 1$, where α is a constant that remains unchanged as I_j increases, and $\lfloor\cdot\rfloor$ denotes the floor function. For an index set $\mathcal{B}$, $|\mathcal{B}|$ denotes the number of elements in $\mathcal{B}$.

It is evident that

$$F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,c}) \geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \tag{11}$$

for $h = 1, \dots, d-1$ as the tensor size increases. This holds because the KL divergence is primarily determined by the number of parameters that can be optimized. This implies that, if $|\mathcal{B}_{I_j,m_j}| \gg |\mathcal{B}_h|$, then $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,c}) \geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c})$. Accordingly, if we define the function g_h as

$$g_h(I_1, \dots, I_d) = \frac{|\mathcal{B}_h|}{|\mathcal{B}_{I_j,m_j}|},$$

It is apparent that if one or more elements in the set $\{I_1, \dots, I_d\}$ increase, g_h will monotonically decrease and converge to zero. This can also be interpreted as

$$|\mathcal{B}_{I_j,m_j}| \gg |\mathcal{B}_h|$$

when the tensor size is large enough. Therefore, Equation (11) follows.

Moreover, $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_1,0}) \geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,0})$ always holds because $\mathcal{B}_h \subseteq \mathcal{B}_{h+1}$. Therefore, once we determine the range of c for which $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_1,0})$ holds, we can subsequently establish that

$$\begin{aligned} F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,c}) &\geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \\ &\geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_1,0}) \geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,0}). \end{aligned}$$

[Ghalamkari and Sugiyama, 2021] shows the closed formula of $\mathcal{P}^{\mathcal{B}_1,0}$, which is given as

$$\begin{aligned} \mathcal{P}^{\mathcal{B}_1,0}_{i_1,\dots,i_d} = \prod_{k=1}^{d} \Bigg(\sum_{i'_1=1}^{I_1} \cdots \sum_{i'_{k-1}=1}^{I_{k-1}} \sum_{i'_{k+1}=1}^{I_{k+1}} \\ \cdots \sum_{i'_d=1}^{I_d} \hat{\mathcal{P}}_{i'_1,\dots,i'_{k-1},i_k,i'_{k+1},\dots,i_d} \Bigg). \end{aligned}$$

Thus, $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_1,0})$ can be computed directly, allowing us to estimate its upper bound, $\overline{F}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_1,0})$, as given in Equation (34) in the appendix. Consequently, the range of c that satisfies the inequality $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq \underline{F}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq \overline{F}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_1,0}) \geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_1,0})$ can be easily determined. As we discussed earlier, this range of c also ensures $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,c}) \geq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,0})$. Therefore, the range of c that satisfies $F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,c}) \leq F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,0})$ is simply a subset of the complement of this range. This leads to the following theorem.

Theorem 4. *For many-body approximation of* $\hat{\mathcal{P}}$, *to satisfy the condition* $D_{\mathrm{KL}}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,c}) \leq D_{\mathrm{KL}}(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_h,0})$ *for every* $h = 1, \dots, d$, c *should at least satisfy:*

$$-\min_{j=1,2,\dots,d} l_j \leq c \leq \min_{j=1,2,\dots,d} l_j, \tag{12}$$

where

$$\begin{aligned} l_j &= \frac{2^d\left(\left(s_{min} - 1\right)d + 1\right) I_j \log\left(\tau\right)\left(\tau\right)^{s_{min}}}{(I_j - m_j)\,(I_j - m_j + 1)\prod_{\substack{h=1\\h\neq j}}^{d} (1 + I_h)\left(\tau\right)^{s_{max}}} \\ &< \frac{2^d\left(d + 1\right) I_j \log\left(\tau\right) b^2}{(I_j - m_j)\,(I_j - m_j + 1)\prod_{\substack{h=1\\h\neq j}}^{d} (1 + I_h)\left(a^2\right)}, \\ j &= 1, \dots, d, \quad \tau = \prod_{j=1}^{d} I_j. \end{aligned} \tag{13}$$

The proof of Theorem 4 can be found in Appendix B. From the above theorem, it is easy to observe that $\mathcal{P}^{\mathcal{B}_h,c_0}$ converges to $\mathcal{P}^{\mathcal{B}_h,0}$ as the tensor size increases. Please note that this theorem can also be applied to the high-order Boltzmann machine, where each $I_j = 2$ for $j = 1, \dots, d$.

Moreover, it is worth mentioning that $\mathcal{P}^{\mathcal{B}_h,0}$ has the maximum entropy among $\mathcal{P}^{\mathcal{B}_h,c}$ for all $c \in \mathbb{R}$. This suggests that as the tensor size increases, the m-projection selects the point with the maximum entropy to minimize the KL divergence, as summarized below.

Theorem 5 (Maximum Entropy Principle)**.** *Consider the set:*

$$\widetilde{\mathcal{P}}^{\mathcal{B}} = \bigcup_{c\in\mathbb{R}} \mathcal{P}^{\mathcal{B},c}, \quad \mathcal{P}^{\mathcal{B},c} = \operatorname*{argmin}_{\mathcal{R}\in\mathcal{S}^c_{\mathcal{B}}} D_{\mathrm{KL}}(\hat{\mathcal{P}},\mathcal{R}),$$

we have $\mathcal{P}^{\mathcal{B},0} \in \bigcup_{c\in\mathbb{R}} \mathcal{P}^{\mathcal{B},c}$ and $\mathcal{P}^{\mathcal{B},0}$ maximizes the entropy in the set $\widetilde{\mathcal{P}}^{\mathcal{B}}$.

Please note that, in this theorem, $\mathcal{B}$ can be any index set that satisfies $\mathcal{B} \subseteq \Omega_d^+$, not restricting to $\mathcal{B}_{I_j,m_j}$ or $\mathcal{B}_h$. The proof of Theorem 5 and the definition of the entropy is in Appendix B. This demonstrates that many-body approximation, as a learning model, gradually evolves into a maximum entropy model as the tensor size increases. Moreover, it can be connected to other maximum entropy learning models widely utilized in various machine learning domains [Mezard and Montanari, 2009, Wainwright et al., 2008].

4 SEARCHING ALGORITHM

To verify our theory, we propose an optimization algorithm to search for the c_0 value, which has been mentioned in the previous section. Please note that our method differs from the Legendre decomposition introduced in the preliminaries, particularly in terms of the constraints and optimality conditions.

4.1 OPTIMIZATION PROBLEM

Our optimization problem can be formulated as

$$\begin{aligned} \mathcal{P}^{\mathcal{B},c_0} &= \operatorname*{argmin}_{\mathcal{P}\in\mathcal{S}^{\mathcal{H}}_{\mathcal{B}}} D_{\mathrm{KL}}(\hat{\mathcal{P}},\mathcal{P}), \\ \mathcal{S}^{\mathcal{H}}_{\mathcal{B}} &= \left\{\mathcal{R}\in\mathcal{S} \mid \theta_\alpha = \theta_\beta \text{ for all } \alpha,\beta\in\Omega_d^+\backslash\mathcal{B}\right\}, \end{aligned} \tag{14}$$

This optimization problem (14) can also be recognized as an m-projection onto the e-flat submanifold $\mathcal{S}^{\mathcal{H}}_{\mathcal{B}}$. The resulting distribution of the projection, denoted as $\mathcal{P}^{\mathcal{B},c_0}$, satisfies $\theta_\alpha = c_0$ for all $\alpha \in \Omega_d^+\backslash\mathcal{B}$. According to the principles of information geometry [Amari, 2016], the result of the m-projection to the e-flat submanifold is guaranteed to exist and unique. Obviously, we have the relationship $\mathcal{S}^c_{\mathcal{B}} \subseteq \mathcal{S}^{\mathcal{H}}_{\mathcal{B}}$ for all $c \in \mathbb{R}$. Please note that in our optimization method, $\mathcal{B}$ can become any index set that satisfies $\mathcal{B} \subseteq \Omega_d^+$, not restricting to $\mathcal{B}_{I_j,m_j}$ or $\mathcal{B}_h$.

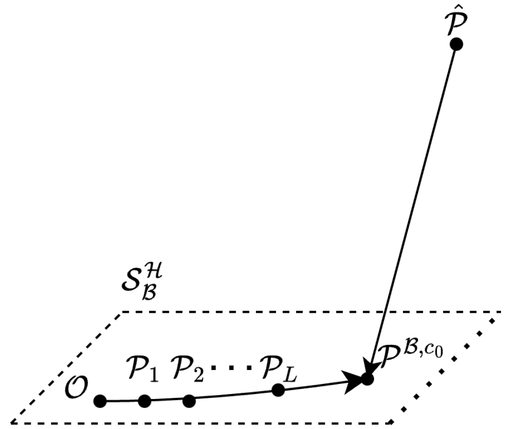

$\hat{\mathcal{P}}$
$\mathcal{S}^{\mathcal{H}}_{\mathcal{B}}$
$\mathcal{P}^{\mathcal{B},c_0}$
$\mathcal{O}$
$\mathcal{P}_1\ \mathcal{P}_2 \cdots \mathcal{P}_L$

Figure 3: Optimization procedure in the submanifold $\mathcal{S}^{\mathcal{H}}_{\mathcal{B}}$. $\hat{\mathcal{P}}$ is an input positive tensor, and $\mathcal{P}^{\mathcal{B},c_0}$ is the result of the optimization problem (14). $\mathcal{P}_L$ is a tensor of the L^{th} step of gradient decent. $\mathcal{O}$ is the initial point of optimization, which is usually a uniform distribution.

4.2 OPTIMIZATION METHOD

First we reformulate the optimization problem as follows:

$$\begin{aligned} &\operatorname*{argmin}_{\mathcal{P}\in\mathcal{S}^{\mathcal{H}}_{\mathcal{B}}} F(\hat{\mathcal{P}},\mathcal{P}) \\ &= -\sum_{i_1=1}^{I_1}\cdots\sum_{i_d=1}^{I_d} \hat{\mathcal{P}}_{i_1,\dots,i_d}\log\mathcal{P}_{i_1,\dots,i_d} \\ &= -\sum_{(i_1,\dots,i_d)\in\Omega_d} \hat{\mathcal{P}}_{i_1,\dots,i_d}\left(\sum_{i'_1=1}^{i_1}\cdots\sum_{i'_d=1}^{i_d}\theta_{i'_1,\dots,i'_d}\right) \end{aligned} \tag{15}$$

subject to $\theta_\alpha = \theta_\beta$ for all $\alpha,\beta \in \Omega_d^+\backslash\mathcal{B}$, where θ_α represents the θ coordinates of the tensor $\mathcal{P} \in \mathcal{S}^{\mathcal{H}}_{\mathcal{B}}$. Since the KL divergence function is convex, and the constraint $\theta_\alpha = \theta_\beta$ is linear, this forms a convex optimization problem. As a result, the optimal solution not only exists but is also unique, aligning with the principles of information geometry that we have previously discussed.

We use the generalized reduced gradient method [Sun and Yuan, 2006] to solve this constrained optimization problem. First, we select an index γ from $\Omega_d^+\backslash\mathcal{B}$ and define $\mathcal{B}_\gamma = \mathcal{B} \cup \{\gamma\}$, $\theta_{\mathcal{B}} = \{\theta_\alpha \mid \alpha \in \mathcal{B}\}$, then we have $\theta_{\Omega_d^+} = \theta_{\Omega_d^+\backslash\mathcal{B}_\gamma} \cup \theta_{\mathcal{B}_\gamma}$. Moreover, from the constraint condition

$$\mathcal{S}^{\mathcal{H}}_{\mathcal{B}} = \left\{Q\in\mathcal{S} \mid \theta_\alpha = \theta_\beta \text{ for all } \alpha,\beta\in\Omega_d^+\backslash\mathcal{B}\right\},$$

we obtain $\theta_\alpha = \theta_\gamma$ for all $\theta_\alpha \in \theta_{\Omega_d^+\backslash\mathcal{B}_\gamma}$. Therefore, we can rewrite F as: $F(\theta_{\Omega_d^+}) = F(\theta_{\Omega_d^+\backslash\mathcal{B}_\gamma}, \theta_{\mathcal{B}_\gamma}) = \widetilde{F}(\theta_{\mathcal{B}_\gamma})$. The number of parameters we need to optimize is $|\mathcal{B}_\gamma| = |\mathcal{B}| + 1$. The gradient of θ_w for each $w \in \mathcal{B}$ is calculated as:

$$\frac{\partial}{\partial \theta_w} \widetilde{F} = \eta_w - \hat{\eta}_w, \tag{16}$$

which is the same as that of Legendre decomposition in Equation (6). The gradient of θ_γ is calculated as

$$\frac{\partial}{\partial \theta_\gamma} \widetilde{F} = \sum_{s \in \Omega_d^+ \setminus \mathcal{B}_\gamma} \frac{\partial}{\partial \theta_s} \widetilde{F} \cdot \frac{\mathrm{d}\theta_s}{\mathrm{d}\theta_\gamma} + \frac{\partial}{\partial \theta_\gamma} \widetilde{F} = \sum_{s \in \Omega_d^+ \setminus \mathcal{B}} (\eta_s - \hat{\eta}_s). \tag{17}$$

Equations (16) and (17) also show that the function $\widetilde{F}$ is minimized if and only if $\eta_w = \hat{\eta}_w$ for all $w \in \mathcal{B}$ and $\sum_{s \in \Omega_d^+ \setminus \mathcal{B}} (\eta_s - \hat{\eta}_s) = 0$. However, in Legendre decomposition, the optimality condition is only given by $\eta_w = \hat{\eta}_w$ for all $w \in \mathcal{B}$, which implies that $\sum_{s \in \Omega_d^+ \setminus \mathcal{B}} (\eta_s - \hat{\eta}_s) = 0$ can further reduce the KL error.

We show the pseudo-code of the above gradient method in Algorithm 1. The time complexity of each iteration is $O(|\Omega_d||\mathcal{B}_\gamma|)$, as that of computing $\mathcal{P}$ from $(\theta_v)_{v \in \mathcal{B}_\gamma}$ (line 5 in Algorithm 1) is $O(|\Omega_d||\mathcal{B}_\gamma|)$ and computing $(\eta_v)_{v \in \Omega_d}$ from $\mathcal{P}$ (line 6 in Algorithm 1) is $O(|\Omega_d|)$. Thus the total complexity is $O\left(h|\Omega_d||\mathcal{B}_\gamma|^2\right)$ with the number of iterations h until convergence.

Although gradient descent is an efficient approach, we can also use the Newton method (natural gradient descent) [Amari, 1998], a second-order optimization method shown in Algorithm 2, to reduce the number of iterations to gain efficiency. Each element of the Hessian matrix $\widetilde{\mathbf{G}} \in \mathbb{R}^{|\mathcal{B}_\gamma| \times |\mathcal{B}_\gamma|}$ of $\widetilde{F}(\theta_{\mathcal{B}_\gamma})$ is calculated as:

$$\begin{aligned}
\widetilde{\mathbf{G}}_{u,v} &= \frac{\partial^2}{\partial \theta_u \partial \theta_v} \widetilde{F} = \mathbf{G}_{u,v}, \quad u, v \in \mathcal{B}, \\
\widetilde{\mathbf{G}}_{\gamma,v} &= \frac{\partial}{\partial \theta_v} \Big(\sum_{s \in \Omega_d^+ \setminus \mathcal{B}} (\eta_s - \hat{\eta}_s) \Big) = \sum_{s \in \Omega_d^+ \setminus \mathcal{B}} \mathbf{G}_{s,v}, \quad v \in \mathcal{B}, \\
\widetilde{\mathbf{G}}_{v,\gamma} &= \sum_{s \in \Omega_d^+ \setminus \mathcal{B}} \mathbf{G}_{v,s}, \quad v \in \mathcal{B}, \text{ and} \\
\widetilde{\mathbf{G}}_{\gamma,\gamma} &= \frac{\partial}{\partial \theta_\gamma} \Big(\sum_{s \in \Omega_d^+ \setminus \mathcal{B}} (\eta_s - \hat{\eta}_s) \Big) = \sum_{s,t \in \Omega_d^+ \setminus \mathcal{B}} \mathbf{G}_{s,t},
\end{aligned} \tag{18}$$

where $\mathbf{G} = (\mathbf{G}_{u,v}) \in \mathbb{R}^{|\Omega_d^+| \times |\Omega_d^+|}$ is the Hessian matrix of $F(\theta_{\Omega_d^+})$ calculated as

$$\begin{aligned}
\mathbf{G}_{u,v}(\theta) &= \frac{\partial \eta_u}{\partial \theta_v} = \frac{\partial^2 F}{\partial \theta_u \partial \theta_v} \\
&= \sum_{w \in \Omega_d} \zeta(u,w)\zeta(v,w)\mathcal{P}_w - \eta_u \eta_v,
\end{aligned} \tag{19}$$

where $\zeta(u, v) = 1$ if $u \leq v$ and $\zeta(u, v) = 0$ otherwise.

The time complexity of each iteration is $O(|\Omega_d||\mathcal{B}_\gamma| + |\mathcal{B}_\gamma|^3)$, where $O(|\Omega_d||\mathcal{B}_\gamma|)$ is needed to compute $\mathcal{P}$ from θ and $O\left(|\mathcal{B}_\gamma|^3\right)$ to compute the inverse of $\widetilde{\mathbf{G}}$, resulting in the total complexity $O\left(h'|\Omega_d||\mathcal{B}_\gamma| + h'|\mathcal{B}_\gamma|^3\right)$ with the number of iterations h' until convergence. We illustrate the optimization procedure in Figure 3.

Algorithm 1 Gradient Descent Algorithm

```
 1: procedure GradientDescent($\hat{\mathcal{P}}, \mathcal{B}_\gamma$)
 2:     Initialize $(\theta_k)_{k \in \Omega_d^+}$        ▷ e.g., $\theta_k = 0$ for all $k$
 3:     repeat
 4:         for each $t \in \mathcal{B}_\gamma = \{v \mid v \in \mathcal{B}\} \cup \{\gamma\}$ do
 5:             Compute $\mathcal{P}$ using current $(\theta_t)_{t \in \mathcal{B}_\gamma}$
 6:             Update $\eta_k$ for each $k \in \Omega_d^+$ from $\mathcal{P}$
 7:             $\theta_v \leftarrow \theta_v - \epsilon(\eta_v - \hat{\eta}_v), \quad v \in \mathcal{B}$
 8:             $\theta_\gamma \leftarrow \theta_\gamma - \epsilon\left(\sum_{s \in \Omega_d^+ \setminus \mathcal{B}} (\eta_s - \hat{\eta}_s)\right)$
 9:         end for
10:     until convergence of $(\theta_t)_{t \in \mathcal{B}_\gamma}$
11: end procedure
```

Algorithm 2 Natural Gradient Algorithm

```
procedure NaturalGradient($\hat{\mathcal{P}}, \mathcal{B}_\gamma$)
    Initialize $(\theta_k)_{k \in \Omega_d^+}$        ▷ e.g., $\theta_k = 0$ for all $k$
    repeat
        Compute $\mathcal{P}$ using current $(\theta_t)_{t \in \mathcal{B}_\gamma}$
        Update $\eta_k$ for each $k \in \Omega_d^+$ from $\mathcal{P}$
        Compute matrix $\mathbf{G}$ and $\widetilde{\mathbf{G}}$ using $\eta_k, k \in \Omega_d^+$
        Compute
            $\Delta\boldsymbol{\eta} \leftarrow \begin{pmatrix} \eta_v - \hat{\eta}_v \\ \sum_{s \in \Omega_d^+ \setminus \mathcal{B}} (\eta_s - \hat{\eta}_s) \end{pmatrix}, \quad v \in \mathcal{B}$
        Invert matrix $\widetilde{\mathbf{G}}$ to get $\widetilde{\mathbf{G}}^{-1}$
        $\theta \leftarrow \theta - \epsilon \widetilde{\mathbf{G}}^{-1} \Delta\boldsymbol{\eta}$
    until convergence of $(\theta_t)_{t \in \mathcal{B}_\gamma}$
end procedure
```

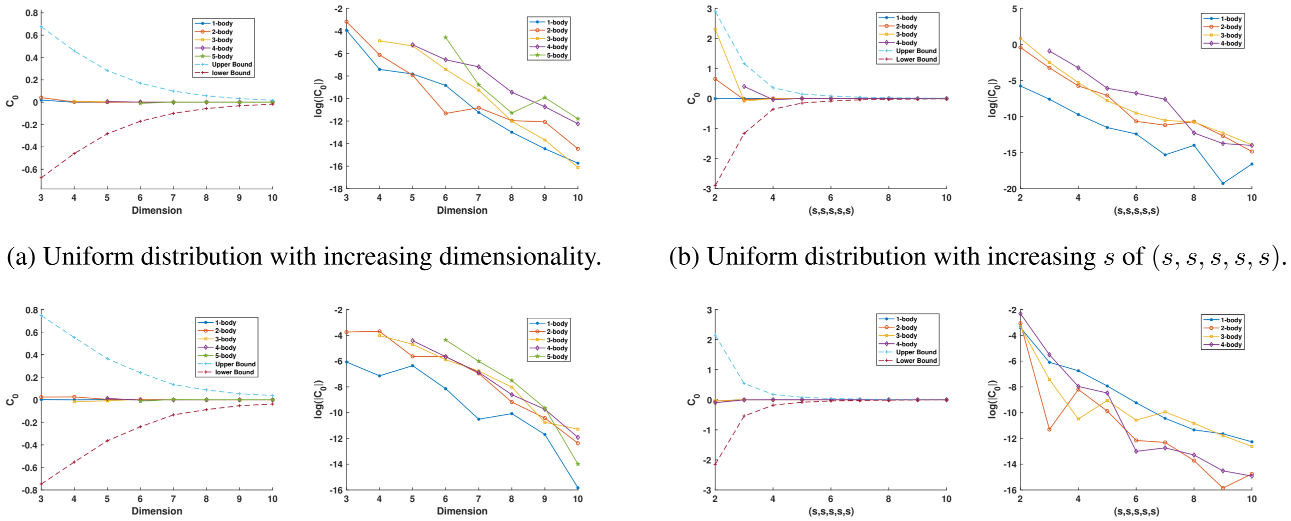
5 NUMERICAL EXPERIMENTS

We numerically examine our theoretical results using synthetic and real-world datasets. Experiments were conducted on Ubuntu 22.04.4 LTS with 88 CPU threads of 2.20GHz Intel Xeon E7-8880 v4 and 3TB of memory.

5.1 EXPERIMENTS SETUP

Synthetic datasets. We generate tensors from the uniform continuous distribution in $[5, 8]$. In experiment $(\mathbf{a})$, we progressively increase the tensor size from $(3, 3, 3)$ to $(3, 3, 3, 3)$, adding one dimension at a time until reaching $(3, 3, 3, 3, 3, 3, 3, 3, 3, 3)$. In experiment $(\mathbf{b})$, we expand the tensor size from $(2, 2, 2, 2, 2)$ to $(3, 3, 3, 3, 3)$, continuing this process until it reaches $(10, 10, 10, 10, 10)$.

Real-world datasets. In the first real data experiment, we



(a) Uniform distribution with increasing dimensionality.

(b) Uniform distribution with increasing s of (s, s, s, s, s).

(c) Butterfly figure with increasing dimensionality.

(d) Butterfly figure with increasing s of (s, s, s, s, s).

Figure 4: Experimental results for uniform distribution and butterfly figure. (a, c) The horizontal axis is the total dimension for the input tensor, with each mode having 3 elements. (b, d) The horizontal axis is the value of s, and the total size of the input tensor is (s, s, s, s, s). The vertical axis shows the value of c_0 or the log value of c_0 in $\mathcal{P}^{\mathcal{B}_h, c_0}$.

utilize the TokyoTech hyperspectral image data set [Monno et al., 2015, 2017]. For each image, it is a $(500, 500, 31)$ tensor, where each mode represents the width, height, and 31-band hyperspectral images from 420 to 720 nm at 10 nm intervals, respectively. We choose the first figure in the dataset, which is a butterfly image, and each pixel value lies within the range $[0.00265, 1]$. In experiment $(\mathbf{c})$, a sub-tensor was extracted from the original tensor, corresponding to the segment $[249{:}330, 249{:}330, 1{:}9]$, and then extracted it and reshaped into a $(3, 3, 3)$ tensor. This sub-tensor was subsequently expanded and reshaped into a $(3, 3, 3, 3)$ tensor, with the process continuing until it reached the final shape of $(3, 3, 3, 3, 3, 3, 3, 3, 3, 3)$.

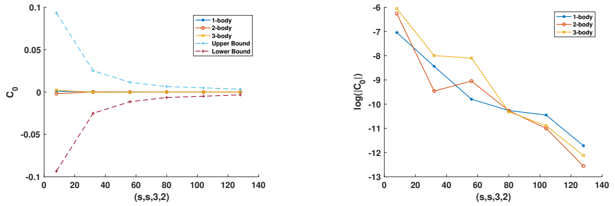
Furthermore, in experiment $(\mathbf{d})$, from the original tensor, we extracted a sub-tensor defined by the segment $[250 : 350, 250 : 350, 1 : 10]$, which was subsequently reshaped into a $(10, 10, 10, 10, 10)$ tensor. To progressively enlarge the tensor, we first extracted a smaller sub-tensor of size $(2, 2, 2, 2, 2)$ and then expanded it to $(3, 3, 3, 3, 3)$. We continued this process incrementally until the tensor reached its final size of $(10, 10, 10, 10, 10)$. In the second real data experiment, we used the Columbia Object Image Library (COIL-100) dataset [Nene et al., 1996]; for each image, it can be regarded as a tensor of size $(128, 128, 3)$. We randomly picked two images and combined them as a $(128, 128, 3, 2)$ tensor, where each mode represents the width, height, color, and image index, respectively. Each pixel value falls within the range $[1, 255]$. We increase the tensor from $(4, 4, 3, 2)$ to $(128, 128, 3, 2)$ in increments of 24 at each step for width and height channels.

5.2 EXPERIMENTS RESULTS

We show the experimental results for the uniform distribution and butterfly figure in Figure 4 and those for the COIL-100 data in Figure 5. We plot both the c_0 and $\log |c_0|$ to clearly show its trends. These results show that as the tensor size increases, $|c_0|$ of $\mathcal{P}^{\mathcal{B}_h, c_0}$ gradually decreases for any many-body structure, and the results remain consistent with our theoretical bounds. We used $m_j = \lfloor I_j / 2 \rfloor$ and the actual values of $s_{\min}$ and $s_{\max}$ of each sub-tensor to compute the theoretical bound. Furthermore, the convergence rate to zero varies depending on the many-body structure. Specifically, lower-body approximations (one or two body) tend to converge faster than higher-body approximations (three, four, or five body). In addition, for these five different experiments, the $|c_0|$ values fall within the range of $[1 \times 10^{-8}, 1 \times 10^{-5}]$ when the dimension or s is increased to 10, indicating that they are close to zero.

6 CONCLUSION

In this paper, we have discussed the hyperparameter c_0 (or the submanifold $S^{c_0}_{\mathcal{B}_h}$) selection problem in many-body approximation in the optimization problem of minimizing the KL-divergence between the original distribution and statistic model. Our theoretical result shows the asymptotic characteristic of the hyperparameter c_0, which means that as the tensor size increases, the value of c_0 converges to 0. The experimental results in the synthetic and real-world datasets validate our theoretical analysis. This paper not only provides a theoretical foundation for the widely used



(a) Without log transformation

(b) With log transformation

Figure 5: Experimental results for COIL-100 dataset. The horizontal axis is the value of s, and the total size of the input tensor is $(s, s, 3, 2)$.

many-body approximation under large-scale parameters but also proposes an optimal many-body model selection algorithm for small-scale non-negative tensors or empirical distributions.

Acknowledgements

This work was supported by JST, CREST Grant Number JPMJCR22D3, Japan.

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2013.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. American Mathematical Soc., 2000.

Kazu Ghalamkari and Mahito Sugiyama. Fast tucker rank reduction for non-negative tensors using mean-field approximation. In *Advances in Neural Information Processing Systems*, volume 34, pages 443–454. Curran Associates, Inc., 2021.

Kazu Ghalamkari, Mahito Sugiyama, and Yoshinobu Kawahara. Many-body approximation for non-negative tensors. In *Advances in Neural Information Processing Systems*, volume 36, pages 74077–74102. Curran Associates, Inc., 2023.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Yusuke Monno, Daisuke Kiku, Masayuki Tanaka, and Masatoshi Okutomi. Adaptive residual interpolation for color and multispectral image demosaicking. *Sensors*, 17 (12):2787, 2017.

Yusukex Monno, Sunao Kikuchi, Masayuki Tanaka, and Masatoshi Okutomi. A practical one-shot multispectral imaging system using a single image sensor. *IEEE Transactions on Image Processing*, 24(10):3048–3059, 2015.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library: COIL-100. Technical Report CUCS-006-96, Department of Computer Science, Columbia University, 1996.

Terrence Sejnowski. Higher-order boltzmann machines. In *AIP Conference Proceedings*, volume 151, pages 398–403. American Institute of Physics, 1986.

Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Sparse nested Markov models with log-linear parameters. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 576–585. AUAI Press, 2013.

Mahito Sugiyama, Hiroyuki Nakahara, and Koji Tsuda. Tensor balancing on statistical manifold. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3270–3279. PMLR, 2017.

Mahito Sugiyama, Hiroyuki Nakahara, and Koji Tsuda. Legendre decomposition for tensors. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Wenyu Sun and Ya-Xiang Yuan. *Optimization theory and methods: nonlinear programming*. Springer Science & Business Media, 2006.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2): 1–305, 2008.

Optimal Submanifold Structure in Log-linear Models (Supplementary Material)

Derun Zhou[1,2] **Mahito Sugiyama**[1,2]

[1]National Institute of Informatics, Tokyo, Japan
[2]The Graduate University for Advanced Studies, SOKENDAI

A PROJECTION THEORY IN INFORMATION GEOMETRY

We explain concepts of information geometry used in this study, including natural parameters, expectation parameters, and model flatness. In the following discussion, we consider only discrete probability distributions.

A.1 (θ, η)-COORDINATE AND GEODESICS

In this study, we regard a normalized d-order non-negative tensor $\mathcal{P} \in \mathbb{R}_{>0}^{I_1 \times \cdots \times I_d}$ as a discrete probability distribution with d random variables and the ith random variable can take values in $\{1, \ldots, I_i\}$. Let $\mathcal{S}$ be the set of discrete probability distributions with d random variables. The entire space $\mathcal{S}$ is a non-Euclidean space, where the Fisher information matrix $\mathbf{G}$ serves as the Riemann metric. This metric arises from the second-order differentiation of the KL divergence, as shown in Equation (19). In Euclidean space, a straight line is the shortest path between two points. In a non-Euclidean space, such a shortest path is called a geodesic. In the space $\mathcal{S}$, two kinds of geodesics can be introduced: e-geodesics and m-geodesics. For two points $\mathcal{P}_1, \mathcal{P}_2 \in \mathcal{S}$, e- and m-geodesics are defined as

$$\{\mathcal{R}_t \mid \log \mathcal{R}_t = (1-t)\log \mathcal{P}_1 + t \log \mathcal{P}_2 - \phi(t)\}, \quad \{\mathcal{R}_t \mid \mathcal{R}_t = (1-t)\mathcal{P}_1 + t\mathcal{P}_2\},$$

respectively, where $0 \leq t \leq 1$ and $\phi(t)$ is a normalization factor to keep $\mathcal{R}_t$ to be a distribution.

We can parameterize the distributions $\mathcal{P} \in \mathcal{S}$ using parameters known as natural parameters. In Equation (1), we have described the relationship between a distribution $\mathcal{P}$ and a natural parameter vector $\boldsymbol{\theta} = (\theta_{2,\ldots,1}, \ldots, \theta_{I_1,\ldots,I_d})$. The natural parameter θ serves as a coordinate system of $\mathcal{S}$, hence any distribution in $\mathcal{S}$ is specified by determining $\boldsymbol{\theta}$. Furthermore, we can also specify a distribution $\mathcal{P}$ by its expectation parameter vector $\boldsymbol{\eta} = (\eta_{2,\ldots,1}, \ldots, \eta_{I_1,\ldots,I_d})$, which corresponds to expected values of the distribution and an alternative coordinate system of $\mathcal{S}$. The definition of the expectation parameter $\boldsymbol{\eta}$ is described in Equations (3) and (4). The pair of coordinates, θ-coordinates and η-coordinates, are orthogonal with each other, which means that the Fisher information matrix $\mathbf{G}$ has the following property, $\mathbf{G}_{u,v} = \partial \eta_u / \partial \theta_v$ and $\left(\mathbf{G}^{-1}\right)_{u,v} = \partial \theta_u / \partial \eta_v$. We can describe e- and m-geodesics using these parameters as follows.

$$\left\{\boldsymbol{\theta}^t \mid \boldsymbol{\theta}^t = (1-t)\boldsymbol{\theta}^{\mathcal{P}_1} + t\boldsymbol{\theta}^{\mathcal{P}_2}\right\}, \quad \left\{\boldsymbol{\eta}^t \mid \boldsymbol{\eta}^t = (1-t)\boldsymbol{\eta}^{\mathcal{P}_1} + t\boldsymbol{\eta}^{\mathcal{P}_2}\right\},$$

where $\boldsymbol{\theta}^{\mathcal{P}}$ and $\boldsymbol{\eta}^{\mathcal{P}}$ are θ- and η-coordinate of a distribution $\mathcal{P} \in \mathcal{S}$, respectively.

A.2 FLATNESS AND PROJECTIONS

A submanifold is called e-flat if any e-geodesic connecting two points in it remains within the submanifold. The vertical descent of an m-geodesic from a point $\mathcal{P} \in \mathcal{S}$ onto an e-flat submanifold $\mathcal{S}_{\mathcal{B}_{e\text{-}flat}}$ is called the m-projection. Similarly, the e-projection is obtained by interchanging e and m. The flatness of subspaces guarantees the uniqueness of the projection destination, denoted as $\mathcal{P}_{e\text{-flat}}$ or $\mathcal{P}_{m\text{-flat}}$, which minimizes the following KL divergence:

$$\mathcal{P}_{e\text{-flat}} = \underset{\mathcal{Q}\in\mathcal{S}_{\mathcal{B}_{e\text{-flat}}}}{\operatorname{argmin}}\; D_{KL}(\mathcal{P},\mathcal{Q}),$$
$$\mathcal{P}_{m\text{-flat}} = \underset{\mathcal{Q}\in\mathcal{S}_{\mathcal{B}_{m\text{-flat}}}}{\operatorname{argmin}}\; D_{KL}(\mathcal{Q},\mathcal{P}).$$

A.3 THEORETICAL REMARKS

A submanifold with some natural parameters fixed at some constant value c is e-flat, which follows directly from the definition of e-flatness. Here, our discussion focuses on m-projection onto the submanifold $\mathcal{S}^c_{\mathcal{B}} = \{Q \in \mathcal{S} \mid \theta_v = c \text{ for all } v \in \Omega^+_d \setminus \mathcal{B}\}$, where $\mathcal{B}$ can be any index set satisfying $\mathcal{B} \subseteq \Omega^+_d$. Since the constraint $\theta_v = c$ is linear and the KL divergence function is convex, the optimal solution $\mathcal{P}^{\mathcal{B},c} = \operatorname{argmin}_{\mathcal{R}\in\mathcal{S}^c_{\mathcal{B}}} D_{KL}(\hat{\mathcal{P}},\mathcal{R})$ always uniquely exists. From another perspective, the e-flat $\mathcal{S}^c_{\mathcal{B}}$ forms a convex set. Consequently, this optimization problem involves minimizing a convex function over a convex set, thereby classified as a convex optimization problem. If a space is both e-flat and m-flat, it is called dually-flat. The space $\mathcal{S}$ of discrete probability distributions is dually-flat.

A.4 EXAMPLES FOR MÖBIUS FUNCTION

In the proposed method, we transform the distribution $\mathcal{P} \in \mathbb{R}^{I_1\times\cdots\times I_d}_{>0}$ using the Möbius function, defined in Section 2.1. By Equation (4), we can express $\mathcal{P}$ in terms of the expectation parameter $\boldsymbol{\eta}$. For example, for $d = 2, 3$:

$$\begin{aligned}
\mathcal{P}_{i_1,i_2} &= \eta_{i_1,i_2} - \eta_{i_1+1,i_2} - \eta_{i_1,i_2+1} + \eta_{i_1+1,i_2+1},\\
\mathcal{P}_{i_1,i_2,i_3} &= \eta_{i_1,i_2,i_3} - \eta_{i_1+1,i_2,i_3} - \eta_{i_1,i_2+1,i_3} - \eta_{i_1,i_2,i_3+1}\\
&\quad + \eta_{i_1+1,i_2+1,i_3} + \eta_{i_1+1,i_2,i_3+1} + \eta_{i_1,i_2+1,i_3+1} - \eta_{i_1+1,i_2+1,i_3+1},
\end{aligned}$$

where we assume $\eta_{I_1+1,i_2} = \eta_{i_1,I_2+1} = 0$ and $\eta_{I_1+1,i_2,i_3} = \eta_{i_1,I_2+1,i_3} = \eta_{i_1,i_2,I_3+1} = 0$.

A.5 EXAMPLES FOR APPLICATION IN TENSOR DECOMPOSITION

An application where the effectiveness of the choice of c can be more easily observed is in the compression of multi-dimensional data (e.g., images). To illustrate this point, we revisit the example introduced earlier. The input tensor is:

$$\begin{bmatrix} 833 & 1 & 2 & 4 & 7 & 4 & 8 \\ 430 & 33 & 5 & 1 & 711 & 112 & 4 \\ 39 & 6 & 29 & 2 & 9 & 3 & 121 \\ 2 & 2 & 8 & 6 & 311 & 10 & 122 \end{bmatrix}.$$

Let us use the model submanifold $\mathcal{S}^c_{\mathcal{B}_1}$ with $\mathcal{B}_1$, which denotes the index set of one-body natural parameters (the first row and column of the θ value in the matrix). θ-parameters of each tensor in the submanifold are in the form of

$$\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} & \theta_{15} & \theta_{16} & \theta_{17} \\ \theta_{21} & c & c & c & c & c & c \\ \theta_{31} & c & c & c & c & c & c \\ \theta_{41} & c & c & c & c & c & c \end{bmatrix}.$$

In the view of multi-dimensional data compression, the original matrix $\mathcal{X}$ requires $4 \times 7 = 28$ values to represent it. However, by approximating it by projecting it onto the submanifold $\mathcal{S}^c_{\mathcal{B}_1}$, we only need to store 11 parameters, the optimized values of the one-body θ-parameters and the constant c (the traditional method just sets $c = 0$, but the number of parameters that need to be stored is still 11). In other words, in our tensor decomposition task, sparsity does not appear in the original matrix space; instead, it manifests in the θ-coordinate space. As shown in the introduction, selecting the submanifold $\mathcal{S}^0_{\mathcal{B}_1}$, results in a KL error of 0.46 and the RMSE of 0.56. In contrast, choosing $\mathcal{S}^{0.54}_{\mathcal{B}_1}$ as the model submanifold reduces the KL error to 0.19 and the RMSE to 0.24—nearly half of the previous values. This demonstrates that varying the value of c can significantly affect the reconstruction quality of the tensor.

A.6 EXAMPLES FOR LOG-LINEAR MODEL

As an example of the *log-linear model* [Agresti, 2013], consider the distribution of an n-dimensional binary vector $\mathbf{x} = (x^1, \dots, x^n) \in \{0,1\}^n$, where the log-probability is expressed as:

$$\log p(\mathbf{x}) = \sum_i \theta^i x^i + \sum_{i<j} \theta^{ij} x^i x^j + \sum_{i<j<k} \theta^{ijk} x^i x^j x^k + \cdots + \theta^{1 \dots n} x^1 x^2 \cdots x^n - \psi,$$

where $\boldsymbol{\theta} = (\theta^1, \dots, \theta^{1 \dots n})$ is the natural parameter vector, and ψ is the log-partition function (normalizer). The corresponding expectation parameters $\boldsymbol{\eta} = (\eta^1, \dots, \eta^{1 \dots n})$ represent the expected values of variable combinations:

$$\eta^i = \mathbb{E}[x^i] = \Pr(x^i = 1), \quad \eta^{ij} = \mathbb{E}[x^i x^j] = \Pr(x^i = x^j = 1), \quad \eta^{1 \dots n} = \mathbb{E}[x^1 \cdots x^n] = \Pr(x^1 = \cdots = x^n = 1).$$

B PROOFS.

Lemma (1). *For any input tensor $\hat{\mathcal{P}} \in \mathbb{R}_{>0}^{I_1 \times \cdots \times I_d}$ and its m-projection $\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}$ onto the submanifold $\mathcal{S}^c_{\mathcal{B}_{I_j,m_j}}$, we have*

$$\sum_{i_j=m_j}^{I_j} \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\dots,i_{j-1},i_j,i_{j+1},\dots,i_d} = \sum_{i_j=m_j}^{I_j} \hat{\mathcal{P}}_{i_1,\dots,i_{j-1},i_j,i_{j+1},\dots,i_d}. \tag{20}$$

Proof. Because $\eta^{\mathcal{B}_{I_j,m_j},c}_{i_1,\dots,i_{j-1},i_j,i_{j+1},\dots,i_d} = \hat{\eta}_{i_1,\dots,i_{j-1},i_j,i_{j+1},\dots,i_d}$, $i_j = 1, \dots, m_j$, and $\mu^{i'_1,\dots,i'_d}_{i_1 \dots i_d} = \prod_{k=1}^d \mu^{i'_k}_{i_k}$, it follows that

$$\begin{aligned}
&\sum_{i_j=m_j}^{I_j} \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\dots,i_{j-1},i_j,i_{j+1},\dots,i_d} \\
&= \sum_{i_j=m_j}^{I_j} \sum_{\left(i'_1,\dots,i'_d\right) \in \Omega_d} \mu^{i'_1,\dots,i'_d}_{i_1 \dots i_d} \eta^{\mathcal{B}_{I_j,m_j},c}_{i'_1,\dots,i'_d} \\
&= \sum_{i_j=m_j}^{I_j} \sum_{\left(i'_1,\dots,i'_d\right) \in \Omega_d} \left(\prod_{k=1}^{d} \mu^{i'_k}_{i_k} \right) \eta^{\mathcal{B}_{I_j,m_j},c}_{i'_1,\dots,i'_d} \\
&= \sum_{i_j=m_j}^{I_j} \sum_{i'_1=i_1}^{i_1+1} \cdots \sum_{i'_{j-1}=i_{j-1}}^{i_{j-1}+1} \sum_{i'_j=i_j}^{i_j+1} \cdots \sum_{i'_d=i_d}^{i_d+1} \left(\prod_{\substack{k=1 \\ k \neq j}}^{d} \mu^{i'_k}_{i_k} \right) \left(\sum_{i'_j=i_j}^{i_j+1} \mu^{i'_j}_{i_j} \eta^{\mathcal{B}_{I_j,m_j},c}_{i'_1,\dots,i'_d} \right) \\
&= \sum_{i'_1=i_1}^{i_1+1} \cdots \sum_{i'_{j-1}=i_{j-1}}^{i_{j-1}+1} \sum_{i'_j=i_j}^{i_j+1} \cdots \sum_{i'_d=i_d}^{i_d+1} \left(\prod_{\substack{k=1 \\ k \neq j}}^{d} \mu^{i'_k}_{i_k} \right) \left(\sum_{i_j=m_j}^{I_j} \left(\eta^{\mathcal{B}_{I_j,m_j},c}_{i'_1,\dots,i'_{j-1},i_j,i'_{j+1},\dots,i'_d} - \eta^{\mathcal{B}_{I_j,m_j},c}_{i'_1,\dots,i'_{j-1},i_j+1,i'_{j+1},\dots,i'_d} \right) \right) \\
&= \sum_{i'_1=i_1}^{i_1+1} \cdots \sum_{i'_{j-1}=i_{j-1}}^{i_{j-1}+1} \sum_{i'_j=i_j}^{i_j+1} \cdots \sum_{i'_d=i_d}^{i_d+1} \left(\prod_{\substack{k=1 \\ k \neq j}}^{d} \mu^{i'_k}_{i_k} \right) \left(\eta^{\mathcal{B}_{I_j,m_j},c}_{i'_1,\dots,i'_{j-1},m_j,i'_{j+1},\dots,i'_d} \right) \\
&= \sum_{i'_1=i_1}^{i_1+1} \cdots \sum_{i'_{j-1}=i_{j-1}}^{i_{j-1}+1} \sum_{i'_j=i_j}^{i_j+1} \cdots \sum_{i'_d=i_d}^{i_d+1} \left(\prod_{\substack{k=1 \\ k \neq j}}^{d} \mu^{i'_k}_{i_k} \right) \left(\hat{\eta}_{i'_1,\dots,i'_{j-1},m_j,i'_{j+1},\dots,i'_d} \right) \\
&= \sum_{i_j=m_j}^{I_j} \hat{\mathcal{P}}_{i_1,\dots,i_{j-1},i_j,i_{j+1},\dots,i_d}.
\end{aligned}$$

□

Theorem (1). *For any input tensor $\hat{\mathcal{P}} \in \mathbb{R}_{>0}^{I_1\times\cdots\times I_d}$, its m-projection $\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}$ onto the submanifold $\mathcal{S}^c_{\mathcal{B}_{I_j,m_j}}$ is given as*

$$\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d} = \hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d}$$

for $i_j = 1, 2, \ldots, m_j - 1$ and

$$\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d} = \left(\sum_{k=0}^{I_j-m_j} \hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d}\right) \exp\left(ck\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right) \left(\sum_{k=0}^{I_j-m_j} \exp\left(ck\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)\right)^{-1}$$

for $k = 0, \ldots, I_j - m_j$.

Proof. Remind that $\eta^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d} = \hat{\eta}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d}, i_j = 1, \ldots, m_j$. We have

$$\begin{aligned} \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d} &= \sum_{i'_1=i_1}^{i_1+1}\cdots\sum_{i'_j=i_j}^{i_j+1}\cdots\sum_{i'_d=i_d}^{i_d+1}\left(\prod_{k=1}^{d}\mu_{i_k}^{i'_k}\right)\eta^{\mathcal{B}_{I_j,m_j},c}_{i'_1,\ldots,i'_d} \\ &= \sum_{i'_1=i_1}^{i_1+1}\cdots\sum_{i'_j=i_j}^{i_j+1}\cdots\sum_{i'_d=i_d}^{i_d+1}\left(\prod_{k=1}^{d}\mu_{i_k}^{i'_k}\right)\hat{\eta}_{i'_1,\ldots,i'_d} \\ &= \hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},i_j,i_{j+1},\ldots,i_d}, i_j = 1, 2, \ldots, m_j - 1. \end{aligned} \tag{21}$$

Moreover,

$$\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d} = \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j,i_{j+1},\ldots,i_d} e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}. \tag{22}$$

Therefore,

$$\begin{aligned} \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j,i_{j+1},\ldots,i_d}\left(\sum_{k=0}^{I_j-m_j} e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}\right) &= \left(\sum_{k=0}^{I_j-m_j}\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d}\right) \\ &= \left(\sum_{k=0}^{I_j-m_j}\hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d}\right), \end{aligned}$$

$$\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j,i_{j+1},\ldots,i_d} = \left(\sum_{k=0}^{I_j-m_j}\hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d}\right)\frac{1}{\left(\sum_{k=0}^{I_j-m_j} e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}\right)}.$$

□

Theorem (2). *$F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq \max\{l, -l\}$, where*

$$l = \frac{c(I_j - m_j)(I_j - m_j + 1)\prod_{\substack{h=1\\h\neq j}}^{d} I_h(1 + I_h)}{2^d\left(\prod_{j=1}^{d} I_j\right)^{s_{min}}}.$$

Proof. Let us consider

$$F(\hat{\mathcal{P}}; \mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) = -\sum_{i_1=1}^{I_1}\cdots\sum_{i_d=1}^{I_d}\left\{\hat{\mathcal{P}}_{i_1\ldots i_d}\log\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_d}\right\}. \tag{23}$$

Because $\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_j,\ldots,i_d} = \hat{\mathcal{P}}_{i_1,\ldots,i_j,\ldots,i_d}$ for every $i_j = 1, 2, \ldots, m_j - 1$ and for any c, we only need to consider

$$f(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) = -\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{k=0}^{I_j-m_j}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left\{\hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d}\log\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d}\right\}. \tag{24}$$

Moreover, $F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq f(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c})$ always hold. On the one hand,

$$\begin{aligned}
\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d} &= \left(\sum_{k=0}^{I_j-m_j}\hat{\mathcal{P}}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d}\right)\frac{e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}}{\left(\sum_{k=0}^{I_j-m_j} e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}\right)} \\
&\leq \frac{e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}}{\left(\sum_{k=0}^{I_j-m_j} e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}\right)} = \frac{1}{\left(\sum_{h=-k}^{I_j-m_j-k} e^{c\left(h\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}\right)} \\
&\leq e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}.
\end{aligned} \tag{25}$$

Therefore, by applying $\min(\hat{\mathcal{P}}) = \frac{1}{\left(\prod_{j=1}^{d} I_j\right)^{s_{min}}}$,

$$\begin{aligned}
f(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) &\geq -\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{k=0}^{I_j-m_j}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left\{\frac{1}{\left(\prod_{j=1}^{d}(I_j)\right)^{s_{min}}}\log\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d}\right\} \\
&\geq -\frac{1}{\left(\prod_{j=1}^{d}(I_j)\right)^{s_{min}}}\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{k=0}^{I_j-m_j}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left\{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)\right\} \\
&= -\frac{c}{\left(\prod_{j=1}^{d}(I_j)\right)^{s_{min}}}\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left(\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)\left(\sum_{k=0}^{I_j-m_j} k\right) \\
&= -\frac{c\,(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d} I_h\,(1+I_h)}{2^d\left(\prod_{j=1}^{d} I_j\right)^{s_{min}}}.
\end{aligned} \tag{26}$$

On the other hand,

$$\begin{aligned}
\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\ldots,i_{j-1},m_j+k,i_{j+1},\ldots,i_d} &\leq \frac{e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}}{\left(\sum_{k=0}^{I_j-m_j} e^{c\left(k\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}\right)} = \frac{1}{\left(\sum_{h=-k}^{I_j-m_j-k} e^{c\left(h\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}\right)} \\
&\leq e^{c\left(-(I_j-m_j-k)\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)}.
\end{aligned} \tag{27}$$

Therefore we have

$$\begin{aligned} f(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) &\geq -\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{k=0}^{I_j-m_j}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left\{\frac{1}{\left(\prod_{j=1}^{d}(I_j)\right)^{s_{min}}}\log\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}_{i_1,\dots,i_{j-1},m_j+k,i_{j+1},\dots,i_d}\right\} \\ &\geq -\frac{1}{\left(\prod_{j=1}^{d}(I_j)\right)^{s_{min}}}\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{k=0}^{I_j-m_j}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left\{c\left(-(I_j-m_j-k)\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)\right\} \\ &= \frac{c}{\left(\prod_{j=1}^{d}(I_j)\right)^{s_{min}}}\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left(\prod_{\substack{s=1\\s\neq j}}^{d} i_s\right)\left(\sum_{k=0}^{I_j-m_j} I_j-m_j-k\right) \\ &= \frac{c\,(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d} I_h\,(1+I_h)}{2^d\left(\prod_{j=1}^{d}(I_j)\right)^{s_{min}}}. \end{aligned} \tag{28}$$

□

Corollary (3)**.** *To satisfy the condition $D_{\mathrm{KL}}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \leq D_{\mathrm{KL}}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},0})$, for every $j = 1,\dots,d$, c should at least satisfy $-l \leq c \leq l$, where*

$$l = \frac{2^d s_{max}\log\left(\prod_{j=1}^{d} I_j\right)}{(I_j-m_j)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\left(\prod_{j=1}^{d} I_j\right)^{s_{max}-s_{min}}} < \frac{2^d\log\left(\prod_{j=1}^{d} I_j\right)b^2}{(I_j-m_j)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\,a^2}. \tag{29}$$

Proof. For $c = 0$, let us define

$$h_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_d} := \sum_{k=0}^{I_j-m_j}\hat{\mathcal{P}}_{i_1,\dots,i_{j-1},m_j+k,i_{j+1},\dots,i_d} \leq \frac{I_j-m_j+1}{\left(\prod_{j=1}^{d}(I_j)\right)^{s_{max}}}.$$

Then it follows that

$$\begin{aligned} \mathcal{P}^{\mathcal{B}_{I_j,m_j},0}_{i_1,\dots,i_{j-1},m_j+k,i_{j+1},\dots,i_d} &= \left(\sum_{k=0}^{I_j-m_j}\hat{\mathcal{P}}_{i_1,\dots,i_{j-1},m_j+k,i_{j+1},\dots,i_d}\right)\frac{1}{I_j-m_j+1} \\ &= \frac{h_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_d}}{I_j-m_j+1}, \end{aligned}$$

$$\begin{aligned} f(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},0}) &= -\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{k=0}^{I_j-m_j}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left\{\hat{\mathcal{P}}_{i_1,\dots,i_{j-1},m_j+k,i_{j+1},\dots,i_d}\log\mathcal{P}^{\mathcal{B}_{I_j,m_j},0}_{i_1,\dots,i_{j-1},m_j+k,i_{j+1},\dots,i_d}\right\} \\ &= -\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{k=0}^{I_j-m_j}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left\{\hat{\mathcal{P}}_{i_1,\dots,i_{j-1},m_j+k,i_{j+1},\dots,i_d}\log\frac{h_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_d}}{I_j-m_j+1}\right\} \\ &= -\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left(\log\frac{h_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_d}}{I_j-m_j+1}\right)\left(\sum_{k=0}^{I_j-m_j}\hat{\mathcal{P}}_{i_1,\dots,i_{j-1},m_j+k,i_{j+1},\dots,i_d}\right) \\ &= -\sum_{i_1=1}^{I_1}\cdots\sum_{i_{j-1}=1}^{I_{j-1}}\sum_{i_{j+1}=1}^{I_{j+1}}\cdots\sum_{i_d=1}^{I_d}\left(\log h_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_d}-\log(I_j-m_j+1)\right)h_{i_1,\dots,i_{j-1},i_{j+1},\dots,i_d} \\ &\leq \frac{s_{max}\log\left(\prod_{j=1}^{d} I_j\right)(I_j-m_j+1)}{\left(\prod_{j=1}^{d}(I_j)\right)^{s_{max}}}\prod_{\substack{h=1\\h\neq j}}^{d} I_h. \end{aligned} \tag{30}$$

If $\underline{f}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq \overline{f}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},0})$, which means

$$c \leq -\frac{2^d s_{max} \log\left(\prod_{j=1}^{d} I_j\right)}{(I_j - m_j)\left(\prod_{j=1}^{d} I_j\right)^{s_{max}-s_{min}} \prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)}$$

or

$$c \geq \frac{2^d s_{max} \log\left(\prod_{j=1}^{d} I_j\right)}{(I_j - m_j)\left(\prod_{j=1}^{d} I_j\right)^{s_{max}-s_{min}} \prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)},$$

then $D_{\mathrm{KL}}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq D_{\mathrm{KL}}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},0})$.
Therefore, c should at least satisfies:

$$-\frac{2^d s_{max} \log\left(\prod_{j=1}^{d} I_j\right)}{(I_j - m_j)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\left(\prod_{j=1}^{d} I_j\right)^{s_{max}-s_{min}}} \leq c \leq \frac{2^d s_{max} \log\left(\prod_{j=1}^{d} I_j\right)}{(I_j - m_j)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\left(\prod_{j=1}^{d} I_j\right)^{s_{max}-s_{min}}}. \tag{31}$$

□

Theorem (4). *For many-body approximation of $\hat{\mathcal{P}}$, to satisfy the condition $D_{\mathrm{KL}}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_h,c}) \leq D_{\mathrm{KL}}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_h,0})$ for every $h = 1,\ldots,d$, c should at least satisfy:*

$$-\min_{j=1,2,\ldots,d} l_j \leq c \leq \min_{j=1,2,\ldots,d} l_j, \tag{32}$$

where

$$l_j = \frac{2^d\left((s_{min}-1)\,d+1\right) I_j \log(\tau)\,(\tau)^{s_{min}}}{(I_j - m_j)(I_j - m_j + 1)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\,(\tau)^{s_{max}}} < \frac{2^d\,(d+1)\, I_j \log(\tau)\, b^2}{(I_j - m_j)(I_j - m_j + 1)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\,(a^2)},$$

$$j = 1,\ldots,d, \quad \tau = \prod_{j=1}^{d} I_j.$$

Proof. According to the closed formula of $\mathcal{P}^{\mathcal{B}_1,0}$,

$$\begin{aligned}
\mathcal{P}^{\mathcal{B}_1,0}_{i_1,\ldots,i_d} &= \prod_{k=1}^{d}\left(\sum_{i'_1=1}^{I_1}\cdots\sum_{i'_{k-1}=1}^{I_{k-1}}\sum_{i'_{k+1}=1}^{I_{k+1}}\cdots\sum_{i'_d=1}^{I_d}\hat{\mathcal{P}}_{i'_1,\ldots,i'_{k-1},i_k,i'_{k+1},\ldots,i_d}\right)\\
&\geq \prod_{k=1}^{d}\left(\frac{1}{\left(\prod_{j=1}^{d} I_j\right)^{s_{min}}}\prod_{\substack{j=1\\j\neq k}}^{d} I_j\right)\\
&= \left(\prod_{j=1}^{d} I_j\right)^{(1-s_{min})d-1}.
\end{aligned} \tag{33}$$

Moreover, $D_{\mathrm{KL}}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq D_{\mathrm{KL}}(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_1,0})$ if and only if $F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_1,0})$. Therefore,

$$\begin{aligned}
F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_1,0}) &= -\sum_{i_1=1}^{I_1}\cdots\sum_{i_d=1}^{I_d}\hat{\mathcal{P}}_{i_1\ldots i_d}\log\mathcal{P}^{\mathcal{B}_1,0}_{i_1,\ldots,i_d}\\
&\leq -\sum_{i_1=1}^{I_1}\cdots\sum_{i_d=1}^{I_d}\frac{1}{\left(\prod_{j=1}^{d} I_j\right)^{s_{max}}}\log\mathcal{P}^{\mathcal{B}_1,0}_{i_1,\ldots,i_d}\\
&\leq \frac{\left((s_{min}-1)\,d+1\right)\left(\prod_{j=1}^{d} I_j\right)\log\prod_{j=1}^{d} I_j}{\left(\prod_{j=1}^{d} I_j\right)^{s_{max}}}.
\end{aligned} \tag{34}$$

On the one hand,

$$\begin{aligned} F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) &\geq f(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \\ &\geq \frac{c\,(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d} I_h\,(1+I_h)}{2^d\left(\prod_{j=1}^{d} I_j\right)^{s_{min}}}. \end{aligned} \tag{35}$$

On the other hand,

$$\begin{aligned} F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) &\geq f(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \\ &\geq -\frac{c\,(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d} I_h\,(1+I_h)}{2^d\left(\prod_{j=1}^{d} I_j\right)^{s_{min}}}. \end{aligned} \tag{36}$$

If

$$\begin{aligned} &-\frac{c\,(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d} I_h\,(1+I_h)}{2^d\left(\prod_{j=1}^{d} I_j\right)^{s_{min}}} \geq \frac{\left((s_{min}-1)\,d+1\right)\left(\prod_{j=1}^{d} I_j\right)\log\prod_{j=1}^{d} I_j}{\left(\prod_{j=1}^{d} I_j\right)^{s_{max}}} \\ &\Leftrightarrow c \leq -\frac{2^d\left((s_{min}-1)\,d+1\right)I_j\log\prod_{j=1}^{d} I_j}{(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\left(\prod_{j=1}^{d} I_j\right)^{s_{max}-s_{min}}} \end{aligned} \tag{37}$$

or

$$\begin{aligned} &\frac{c\,(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d} I_h\,(1+I_h)}{2^d\left(\prod_{j=1}^{d} I_j\right)^{s_{min}}} \geq \frac{\left((s_{min}-1)\,d+1\right)\left(\prod_{j=1}^{d} I_j\right)\log\prod_{j=1}^{d} I_j}{\left(\prod_{j=1}^{d} I_j\right)^{s_{max}}} \\ &\Leftrightarrow c \geq \frac{2^d\left((s_{min}-1)\,d+1\right)I_j\log\prod_{j=1}^{d} I_j}{(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\left(\prod_{j=1}^{d} I_j\right)^{s_{max}-s_{min}}}, \end{aligned} \tag{38}$$

then $F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_h,c}) \geq F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq f(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_{I_j,m_j},c}) \geq F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_1,0}) \geq F(\hat{\mathcal{P}};\mathcal{P}^{\mathcal{B}_h,0})$. Therefore, c should at least satisfies

$$-\min_{j=1,2,\ldots,d} l_j \leq c \leq \min_{j=1,2,\ldots,d} l_j, \tag{39}$$

where

$$\begin{aligned} l_j &= \frac{2^d\left((s_{min}-1)\,d+1\right)I_j\log\left(\prod_{j=1}^{d} I_j\right)\left(\prod_{j=1}^{d} I_j\right)^{s_{min}}}{(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\left(\prod_{j=1}^{d} I_j\right)^{s_{max}}} \\ &< \frac{2^d\,(d+1)\,I_j\log\left(\prod_{j=1}^{d} I_j\right)b^2}{(I_j-m_j)\,(I_j-m_j+1)\prod_{\substack{h=1\\h\neq j}}^{d}(1+I_h)\,a^2} \end{aligned}$$

for each $j = 1,\ldots,d$. □

Theorem (5). *Consider the set:*

$$\widetilde{\mathcal{P}}^{\mathcal{B}} = \bigcup_{c\in\mathbb{R}}\mathcal{P}^{\mathcal{B},c}, \quad \mathcal{P}^{\mathcal{B},c} = \operatorname*{argmin}_{\mathcal{R}\in\mathcal{S}_{\mathcal{B}}^{c}} D_{\mathrm{KL}}(\hat{\mathcal{P}},\mathcal{R}),$$

we have $\mathcal{P}^{\mathcal{B},0} \in \bigcup_{c\in\mathbb{R}}\mathcal{P}^{\mathcal{B},c}$ *and* $\mathcal{P}^{\mathcal{B},0}$ *maximizes the entropy in the set* $\widetilde{\mathcal{P}}^{\mathcal{B}}$*.*

Proof. The Legendre transformation [Amari and Nagaoka, 2000] of $\psi(\theta) = -\theta_{1,\ldots,1}$ is given as

$$\varphi(\eta) = \max_{\theta'}\left(\theta'\eta - \psi(\theta')\right), \quad \theta'\eta = \sum_{x\in\Omega_d^+}\theta'_x\eta_x.$$

Then $\varphi(\eta)$ coincides with the negative entropy, which is defined as

$$\varphi(\eta) = \sum_{(i_1,\dots,i_d)\in\Omega_d} \mathcal{P}_{i_1,\dots,i_d} \log \mathcal{P}_{i_1,\dots,i_d}.$$

Thus, it is clear that $\varphi(\eta)$ is a convex function that attains the minimum value. Moreover, we also have

$$\frac{\partial\varphi(\eta)}{\partial\eta_x} = \frac{\partial}{\partial\eta_x}(\theta\eta - \psi(\theta)) = \theta_x.$$

This holds for all $\mathcal{P}^{\mathcal{B},c}$, $\mathcal{P}^{\mathcal{B},c} = \mathcal{S}_{\mathcal{B}}^{c} \cap \mathcal{S}_{\hat{\mathcal{P}}}^{\mathcal{B}}$. Therefore, $\mathcal{P}^{\mathcal{B},0}$ satisfies the following.

$$\frac{\partial\varphi(\eta)}{\partial\eta_v} = \frac{\partial}{\partial\eta_v}(\theta\eta - \psi(\theta)) = \theta_v = 0, \quad \forall v \in \Omega_d^+ \backslash \mathcal{B},$$

$$\eta_s = \hat{\eta}_s \quad \forall s \in \mathcal{B}.$$

This shows that $\varphi(\eta(\mathcal{P}^{\mathcal{B},0}))$ obtains the minimum value in the set $\bigcup_{c\in\mathbb{R}} \mathcal{P}^{\mathcal{B},c}$. □