Deformed Decomposition for Non-negative Tensors

Kazu Ghalamkari **Petr Taborsky** **Morten Mørup**
Technical University of Denmark

Abstract

Non-negative tensor factorization finds widespread use in numerous applications, however, its global optimization has been a long-standing challenge. In particular, the Frobenius norm minimization, even in the rank-1 setting, is an NP-hard problem. We presently reformulate tensor decompositions using deformed algebra, which is associated with a generalized product such that the exponential law holds for generalized exponential functions, and show that the best rank-1 approximation thereby reduces to a convex optimization problem for the rich χ-divergence family. Building on this foundation, we propose the deformed many-body approximation for non-negative tensors, which expands model capacity while maintaining global optimality by preserving the flatness of the model manifold. Introducing latent variables, for a subclass of χ-divergences, we further develop an Expectation-Maximization-based framework for the deformed extension of traditional low-rank approximations as iterative convex subproblems. Through experiments on tensor-based probability mass function estimation, we show that the deformed decompositions provide implicit regularization and robustness against noise and mislabeled data. Beyond ordinary tensor algebra, our findings provide a factorization framework that enables us to leverage various divergences with convex rank-1 and many-body approximations.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Tensors serve as fundamental objects in various fields, including machine learning and statistics, where real-world data are often naturally stored in tensor form (Grasedyck et al., 2013). Low-rank tensor decompositions approximate the given tensor $\mathbf{T}$ by a rank-R tensor $\mathbf{P}$, which is a linear combination of R rank-1 tensors that can be expressed as the outer product of vectors. The approximation typically aims to minimize the Frobenius norm error $||\mathbf{T}-\mathbf{P}||_F$, where the rank R controls the model capacity (Kolda and Bader, 2009). The resulting low-rank tensor $\mathbf{P}$ can be useful for various applications, including regression (Kossaifi et al., 2020), pattern discovery (Mørup, 2011), missing data estimation (Acar et al., 2011), and density estimation (Novikov et al., 2021).

However, despite the usefulness of factorized representations, tensor decompositions pose serious computational challenges. Hillar and Lim (2013) proved in their seminal paper titled "Most tensor problems are NP-hard" that tensor decompositions minimizing the Frobenius norm become NP-hard even in the rank-1 setting. These results have motivated the community to explore alternative factorizations based on divergences other than the Frobenius norm. A series of studies developed KL-divergence-based non-negative factorization, which enforces non-negativity of the factorized representation that promotes parts-based representations (Lee and Seung, 1999; Chi and Kolda, 2012; Hansen et al., 2015). These works are applicable in numerous fields where data are inherently non-negative (Kargas et al., 2018; Glasser et al., 2019; Ibrahim and Fu, 2021).

Notably, Huang and Sidiropoulos (2017) have shown that the best rank-1 approximation minimizing the KL divergence is a convex optimization problem. This result led to two research streams: many-body approximations and Expectation-Maximization (EM)-based low-rank approximations.

In the former stream, information geometric analysis reveals that the flatness of the set of rank-1 tensors is key to guaranteeing the convexity of the optimization (Sugiyama et al., 2018). Based on this insight, the

many-body approximation approaches tensor factorization via an energy function describing the interactions among tensor modes, and ignores higher-order interactions. This modeling preserves the flatness of the model manifold, thereby guaranteeing the global optimum of the KL-divergence (Ghalamkari et al., 2023). The many-body approximation is regarded as a generalized rank-1 approximation because it recovers the ordinary rank-1 approximation when only first-order interactions are considered (Figure 1**a**). By introducing latent variables in the many-body approximation, the model becomes a conventional rank-R ($R \geq 2$) approximation, which destroys the flatness of the model manifold and thus requires non-convex optimization (Ghalamkari et al., 2024, 2025). The EM-based latter stream surrogates the rank-R approximation by invoking a suitably designed E-step in order to relax the M-step into a convex rank-1 approximation (Yeredor and Haardt, 2019; Chege et al., 2022; Flores et al., 2023; Chege et al., 2023). Information geometry describes the monotonic convergence of these algorithms by regarding EM-based optimizations as a variant of the *em*-algorithm that iteratively performs convex *e*-steps and *m*-steps as projections onto flat manifolds (Hino et al., 2024). However, the entire EM- and *em*-procedure remain non-convex.

While the above progress makes the KL-divergence-based factorization tractable, it has also been reported that the KL divergence is not a universally suitable objective (Hitomi and Ohzeki, 2025), as its mode-covering property often leads to overfitting to noise or outliers (Bishop and Nasrabadi, 2006). Machine learning practice often requires various divergences depending on the nature of tasks and objectives. One emerging example is the Tsallis divergence, whose hyperparameter controls the degree of mode-seeking versus mass-covering behavior (Hernandez-Lobato et al., 2016), and its usefulness has been recognized across a variety of applications (Li and Turner, 2016). Other examples include the reverse KL divergence used in expectation propagation (Minka, 2001) and the Jensen-Shannon divergence employed in recent Generative Adversarial Network training methods (Goodfellow et al., 2014). These demands have motivated the development of various divergence-based low-rank approximations (Sra and Dhillon, 2005; Cichocki et al., 2006, 2011); however, such straightforward extensions still remain a non-convex optimization problem even in the rank-1 settings.

Given the current non-convexity issues in optimizing various divergences and the success of tractable KL-divergence optimization via rank-1 approximation, the following questions naturally arise:

Research questions. *Beyond the KL divergence, can the best rank-1 non-negative approximation be obtained for other divergences? If so, is it possible to formulate many-body approximations that globally optimize such divergences, and to design stable em-based algorithms for rank-R approximations ($R \geq 2$) via an e-step relaxing the m-step into a convex rank-1 approximation?*

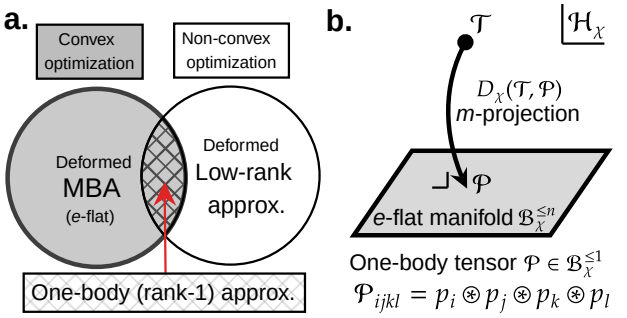


Figure 1: (**a**) The intersection of many-body approximation and low-rank approximation is the one-body (rank-1) approximation. (**b**) One-body approximation of given tensor $\mathbf{T}$ as an orthogonal projection onto a flat manifold in the χ-exponential family $\mathscr{H}_\chi$.

We believe these questions are significant because they lay the foundation for a unified stable optimization framework of non-negative tensor factorizations under various divergences beyond the KL-divergence, allowing us to explore advantages offered by alternative divergences. To address these questions, we provide an analysis based on deformed information geometry (Amari et al., 2012), which is the geometry of the generalized exponential family. We expect it to offer a unified analysis of richer divergences than the standard information geometry associated with only the KL-divergence. As a typical example of a generalized exponential family, q-exponential family defined with q-exponential function $\mathrm{Exp}_q[t] = (1 + (1-q)t)^{1/(1-q)}$ is associated with the robust Tsallis divergence (Tsallis, 1999), and its convex optimization is formulated as a projection onto a flat subspace. We map non-negative normalized tensors as members of such generalized exponential families and apply projection theory in deformed information geometry to develop stable tensor approximations optimizing a rich class of χ-divergences, including the Tsallis divergence and Kaniadakis divergences (Kaniadakis et al., 2005). Moreover, by exploiting deformed algebra, which introduces the deformed product $\circledast$ to extend exponential laws under generalized exponentials, e.g., $\mathrm{Exp}_q[x+y] = \mathrm{Exp}_q[x] \circledast \mathrm{Exp}_q[y]$, we establish the following results:

Our conclusions. *By modifying the definition of outer product in factorized representation, we can formulate a deformed rank-1 and deformed many-body approximations that globally optimize any χ-divergence. In addition, introducing a latent variable in the model, we can formulate a deformed extension of the em-based rank-R approximation ($R \geq 2$) via a convex deformed rank-1 approximation and a closed-form e-step when*

the χ-divergence optimization is equivalent to an f-divergence optimization.

Our novelty lies in simultaneously incorporating deformed algebra and deformed information geometry into tensor decomposition, thereby enabling stable optimization under a variety of divergences, whereas existing generalized-product-based non-negative matrix factorizations lack a guarantee of convexity (Fujimoto and Murata, 2012).

Our experiments on tensor-based probability mass function (PMF) estimation tasks demonstrate that the deformed decompositions (i) provide noise robustness and prevent overfitting, and (ii) the globally optimal deformed many-body approximation achieves comparable performance to baseline methods based on non-convex low-rank decompositions.

2 PRELIMINARIES

We focus on the χ-divergence to develop a unified optimization framework for tensors, as this family includes a broad spectrum of divergences. Within the deformed information geometry associated with χ-divergence, each non-negative tensor is regarded as a member of a generalized exponential family, and the flatness of the set of tensors leads to convex stable optimization (Figure 1**b**). Furthermore, the statistics with generalized exponential law, named deformed statistics, enable tensors to be factorized into a product form, beyond the ordinary modeling based on the exponential family. Accordingly, Section 2.1 reviews the generalized exponential family and its geometry, while Section 2.2 covers deformed statistics. Formal definitions of certain terminology can be found in Appendix D.

2.1 Deformed information geometry

We review extensions of the exponential and logarithmic functions to generalize the exponential family (Naudts, 2002). For any arbitrary positive and increasing differentiable function $\chi : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, we define the χ-logarithm and its inverse, χ-exponential function, as follows:

$$\mathrm{Log}_\chi\,[s] = \int_1^s \frac{dt}{\chi(t)}, \quad \mathrm{Exp}_\chi\,[t] = 1 + \int_0^t \lambda(s)ds, \quad (1)$$

where the function λ is defined such that $\lambda(\mathrm{Log}_\chi\,[s]) = \chi(s)$. Let $\boldsymbol{x}$ be a random variable and Ω be a sample space. For a tuple of N real-valued functions $\boldsymbol{F}(\boldsymbol{x}) = (F_1(\boldsymbol{x}), \dots, F_N(\boldsymbol{x}))$ on the sample space Ω, i.e., $F_n : \Omega \to \mathbb{R}$, the family of distributions $p(\boldsymbol{x}) = \mathrm{Exp}_\chi\left[\boldsymbol{F}(\boldsymbol{x})\boldsymbol{\theta} - \psi_\chi(\boldsymbol{\theta})\right]$ is referred to as the χ-exponential family $\mathscr{H}_\chi$, where each distribution p is parameterized by real-valued parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)^\top$, called the natural parameters. The normalizing constant $\psi_\chi(\boldsymbol{\theta})$, referred to as the χ-free energy, is convex with respect to $\boldsymbol{\theta}$ (Naudts, 2011; Amari, 2021). With an appropriately defined function $\chi(\cdot)$, the χ-exponential families include the q-exponential family (Tsallis, 1988), the (c,d)-exponential family (Hanel and Thurner, 2011), the κ-exponential family (Kaniadakis et al., 2005), and the stretched-exponential family (Anteneodo and Plastino, 1999).

The Legendre transformation of $\psi_\chi(\boldsymbol{\theta})$ is given by $\varphi_\chi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^\top\boldsymbol{\eta} - \psi_\chi(\boldsymbol{\theta})\right)$, and is called the χ-entropy. The dual parameter $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)^\top$ is called the expectation parameter, and it is given as $\boldsymbol{\eta} = \boldsymbol{\nabla}_{\boldsymbol{\theta}}\psi_\chi(\boldsymbol{\theta})$ where $\boldsymbol{\nabla}_{\boldsymbol{\theta}} = (\partial_{\theta_1}, \dots, \partial_{\theta_N})^\top$. As we will see later, since the distribution $p \in \mathscr{H}_\chi$ is determined by specifying either natural parameters $\boldsymbol{\theta}$ or expectation parameters $\boldsymbol{\eta}$, they form two coordinate systems in the manifold $\mathscr{H}_\chi$, called the θ-coordinate system and the η-coordinate system, respectively.

The similarity between distributions $p, q \in \mathscr{H}_\chi$ can be measured by the χ-divergence, which is the Bregman divergence generated by the χ-entropy $\varphi(\boldsymbol{\eta})$. Specifically, it is defined as $D_\chi(p,q) = \varphi_\chi(\boldsymbol{\eta}^p) - \varphi_\chi(\boldsymbol{\eta}^q) - \boldsymbol{\nabla}_{\boldsymbol{\eta}}\varphi_\chi(\boldsymbol{\eta}^q)^\top(\boldsymbol{\eta}^p - \boldsymbol{\eta}^q)$, where $\boldsymbol{\eta}^p$ and $\boldsymbol{\eta}^q$ are the expectation parameters of p and q, respectively, and $\boldsymbol{\nabla}_{\boldsymbol{\eta}} = (\partial_{\eta_1}, \dots, \partial_{\eta_N})^\top$. The explicit form of the χ-divergence is introduced in Section 3.

When $\chi(s) = s$, the family $\mathscr{H}_\chi$ reduces to the ordinary exponential family as Equation (1) recovers the standard logarithm and exponential functions. Consequently, the associated χ-divergence coincides with the KL divergence.

Projection theory in information geometry For a given $p \in \mathscr{H}_\chi$ and a manifold $\mathscr{E} \subset \mathscr{H}_\chi$, the distribution $q^* = \arg\min_{q\in\mathscr{E}} D_\chi(p,q)$ is called the m-projection of p onto $\mathscr{E}$ and the distribution $q^* = \arg\min_{q\in\mathscr{E}} D_\chi(q,p)$ is called e-projection of p onto $\mathscr{E}$. When the manifold $\mathscr{E}$ is constrained by a linear condition in the $\theta(\eta)$-coordinate system, it is called an $e(m)$-flat manifold. It is known that $e(m)$-projection onto any $m(e)$-flat manifold is a convex optimization problem (Amari and Ohara, 2011; Amari et al., 2012).

2.2 Deformed statistics and algebra

Statistics based on the χ-exponential family and χ-divergence is referred to as χ-*statistics*. The χ-exponential function $\mathrm{Exp}_\chi\,[\cdot]$ does not satisfy the exponential law, and it prevents factorizing tensors as seen in Section 3. Thus, we utilize the χ-deformed product $\circledast_\chi$ to obtain a generalized exponential law; that is, $\mathrm{Exp}_\chi\,[a+b] = \mathrm{Exp}_\chi\,[a] \circledast_\chi \mathrm{Exp}_\chi\,[b]$. For simplicity of notation, the dependence of $\circledast$ on χ will often be

omitted in what follows.

The Tsallis statistics is a prominent example of χ-statistics where $\chi(s) = s^q$ for the temperature parameter $q > 0$ (Borges, 2004). The corresponding χ-logarithm and exponential function are given as $\mathrm{Log}_\chi[s] = (s^{1-q} - 1)/(1-q)$ and $\mathrm{Exp}_\chi[t] = (1 + (1-q)t)^{1/(1-q)}$, respectively. The family $\mathscr{H}_\chi$ is called the q-exponential family. The Tsallis deformed product is defined as $x \circledast_q y = (x^{1-q} + y^{1-q} - 1)^{1/(1-q)}$ which reduces to the ordinary product xy as $q \to 1$, and to the sum $x + y - 1$ as $q \to 0$. The parameter q controls the sensitivity to data contamination (Giméenez et al., 2022; Qin and Priebe, 2013). The corresponding χ-divergence is called the Tsallis divergence, and its optimization is equivalent to the α-divergence optimization as we discuss in Section 3. Applications of the q-exponential family are increasingly emerging in the machine learning community, including in logistic regression (Amid et al., 2019), policy optimization (Zhu et al., 2025), likelihood estimation (da Silva et al., 2024), and annealing paths (Masrani et al., 2021).

The Kaniadakis statistics is another example of χ-statistics where $\chi(x) = x/\cosh(\kappa \log x)$ for $-1 < \kappa < 1$ and $\kappa \neq 0$. The χ-logarithm and exponential function are given as $\mathrm{Log}_\chi[s] = \sinh(\kappa \log s)/\kappa$ and $\mathrm{Exp}_\chi[t] = (\kappa t + \sqrt{1 + \kappa^2 t^2})^{1/\kappa}$ (Kaniadakis, 2013). The family $\mathscr{H}_\chi$ is called the κ-exponential family. Kaniadakis statistics have been applied in epidemiology (Kaniadakis et al., 2020), compositional data analysis (Pistone and Shoaib, 2023), and finance (Clementi et al., 2007). As $\kappa \to 0$, the family reduces to the ordinary exponential family.

3 DEFORMED TENSOR MODELS

We denote by $[n]$ the set of natural numbers less than or equal to n, i.e., $[n] = \{1, 2, \ldots, n\}$. Let $\mathscr{S}(D)$ be the set of the entire discrete probability distributions $p(\boldsymbol{\omega})$ with D discrete random variables $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_D)$ for the sample space $\Omega = [I_1] \times \cdots \times [I_D]$. Each member $p \in \mathscr{S}(D)$ has one-to-one correspondence to a normalized non-negative tensor $\mathbf{P} \in \mathbb{R}_{\geq 0}^{I_1 \times \cdots \times I_D}$. More specifically, each tensor element $\mathbf{P}_{i_1 \ldots i_D}$ corresponds to $p(\omega_1 = i_1, \ldots, \omega_D = i_D)$. In the following, no distinction is made between a discrete joint distribution and a normalized non-negative tensor. Consequently, the sample space Ω can be identified with the index set of the tensor $\mathbf{P}$. This interpretation makes it possible to apply information geometry to tensor factorizations (Ghalamkari and Sugiyama, 2023). Throughout the paper, tensor modes denote tensor axes (random variables). All theorems and propositions are formally presented in Appendix E with their proofs.

3.1 Deformed many-body approximation

In the following, we introduce the deformed many-body approximation, which factorizes the given tensor by ignoring higher-order (larger-body) interactions among its modes, formulated as a convex optimization problem. The conventional rank-1 approximation is a particular case of the many-body approximation, as seen below.

For two elements $\boldsymbol{i} = (i_1, \ldots, i_D)$ and $\boldsymbol{j} = (j_1, \ldots, j_D)$ in the index set Ω, we say $\boldsymbol{i} \leq \boldsymbol{j}$ if $i_d \leq j_d$ holds for all $d \in [D]$. The least element $\perp := (1, \ldots, 1) \in \Omega$ satisfies $\perp \leq \boldsymbol{w}$ for any $\boldsymbol{w} \in \Omega$. We define a subset of the index set as $\Omega^+_{\leq \boldsymbol{i}} := \{\boldsymbol{v} \in \Omega^+ \mid \boldsymbol{v} \leq \boldsymbol{i}\}$ where $\Omega^+ = \Omega \backslash \{\perp\}$. Suppose that the tensor $\mathbf{P}$ is parameterized by real-valued parameters $\boldsymbol{\theta} = (\theta_{\boldsymbol{v}})_{\boldsymbol{v} \in \Omega^+}$ as

$$\mathbf{P}_{\boldsymbol{i}} = \mathrm{Exp}_\chi\left[-E_{\boldsymbol{i}} - \psi_\chi(\boldsymbol{\theta})\right], \; E_{\boldsymbol{i}} = -\sum\nolimits_{\boldsymbol{j} \in \Omega^+_{\leq \boldsymbol{i}}} \theta_{\boldsymbol{j}}, \quad (2)$$

where the function $E : \Omega \to \mathbb{R}$ is called the energy function and the χ-free energy $\psi_\chi(\boldsymbol{\theta})$ is defined such that the tensor $\mathbf{P}$ is normalized, i.e., $\sum_{\boldsymbol{i} \in \Omega} \mathbf{P}_{\boldsymbol{i}} = 1$. We note that $E_\perp = 0$ since the empty sum is defined as 0 (Graham et al., 1994). For notational simplicity, the dependence of the energy function E on $\boldsymbol{\theta}$ is not explicitly indicated in the following. As stated in Corollary 1 in Appendix E, the entire set of normalized non-negative tensors with this parameterization forms a χ-exponential family.

We refer to $\theta_{\boldsymbol{v}}$ as a χ-deformed n-body parameter if $\boldsymbol{v} \in \Omega^+$ contains n entries other than 1. For example, with $D = 4$, the parameter θ_{1131} is a χ-deformed one-body parameter, and θ_{1231} is a χ-deformed two-body parameter. We represent χ-deformed n-body parameters as follows:

$$\theta^{(k)}_{i_k} = \theta_{1,\ldots,1,i_k,1,\ldots,1}, \quad \theta^{(k,m)}_{i_k,i_m} = \theta_{1,\ldots,1,i_k,1,\ldots,1,i_m,1,\ldots,1},$$

for $n = 1$ and 2, respectively. As examples, we have $\theta^{(3)}_3 = \theta_{1131}$, $\theta^{(14)}_{23} = \theta_{2113}$, and $\theta^{(134)}_{342} = \theta_{3142}$. To describe the interaction among tensor modes, we define the energy function E as

$$E_{i_1,\cdots,i_D} = \sum_{m=1}^{D} E^{(m)}_{i_m} + \sum_{m=1}^{D} \sum_{k=1}^{m-1} E^{(k,m)}_{i_k,i_m} + \sum_{m=1}^{D} \sum_{k=1}^{m-1} \sum_{p=1}^{k-1} E^{(p,k,m)}_{i_p,i_k,i_m} + \cdots + E^{(1,\ldots,D)}_{i_1,\ldots,i_D}, \quad (3)$$

where the n-th term is given by χ-deformed n-body parameters as

$$E^{(l_1,\ldots,l_n)}_{i_{l_1},\ldots,i_{l_n}} = -\sum\nolimits_{i'_{l_1}=2}^{i_{l_1}} \cdots \sum\nolimits_{i'_{l_n}=2}^{i_{l_n}} \theta^{(l_1,\ldots,l_n)}_{i'_{l_1},\ldots,i'_{l_n}}. \quad (4)$$

We assume $l_1 < l_2 < \cdots < l_n$ without loss of generality. The n-th term in Equation (3) represents the n-body

(n-th order) interactions among n modes of the tensor $\mathbf{P}$. If there is a tuple $(i_{l_1}, \dots, i_{l_n}) \in [I_{l_1}] \times \cdots \times [I_{l_n}]$ that satisfies $E^{(l_1,\dots,l_n)}_{i_{l_1}\dots i_{l_n}} \neq 0$, we say that the χ-deformed n-body interaction among modes $(l_1, \dots, l_n)$ exists in the tensor $\mathbf{P}$. For any integer n, we define a (χ, n)-body tensor as a tensor in which all χ-deformed $m(>n)$-body parameters are zero, such that all the $m(>n)$-body interactions in Equation (3) vanish.

Approximating a given tensor $\mathbf{T}$ by an (χ, n)-body tensor is referred to as the (χ, n)-body approximation of the tensor $\mathbf{T}$. Although the χ-exponential function does not follow the exponentiation rules, the deformed algebra enables tensor factorization via the (χ,n)-body approximation. For example, when $D = 4$ and the (χ, n)-body tensor is denoted by $\mathbf{P}^{\leq n}$, we have

$$\mathbf{T}_{i_1\dots i_4} \simeq \mathbf{P}^{\leq 1}_{i_1\dots i_4} = \frac{1}{Z_\chi} \circledast X^{(1)}_{i_1} \circledast X^{(2)}_{i_2} \circledast X^{(3)}_{i_3} \circledast X^{(4)}_{i_4}, \tag{5}$$

$$\mathbf{T}_{i_1\dots i_4} \simeq \mathbf{P}^{\leq 2}_{i_1\dots i_4} = \frac{1}{Z_\chi} \circledast X^{(1,2)}_{i_1 i_2} \circledast X^{(1,3)}_{i_1 i_3} \circledast X^{(1,4)}_{i_1 i_4} \circledast X^{(2,3)}_{i_2 i_3} \circledast X^{(2,4)}_{i_2 i_4} \circledast X^{(3,4)}_{i_3 i_4 l}, \tag{6}$$

where one- and two-body factors are given as

$$X^{(d)}_{i_d} = \mathrm{Exp}_\chi\left[E^{(d)}_{i_d}\right],$$
$$X^{(k,l)}_{i_k i_l} = \mathrm{Exp}_\chi\left[\frac{1}{3}E^{(k)}_{i_k} + E^{(k,l)}_{i_k i_l} + \frac{1}{3}E^{(l)}_{i_l}\right],$$

respectively, and $\circledast$ represents χ-deformed product. The normalizer Z_χ is defined as $1/Z_\chi = \mathrm{Exp}_\chi\left[-\psi_\chi(\boldsymbol{\theta})\right]$.

In the above discussion, we have considered approximations using all interactions up to n-body (n-th order) interactions. More flexibly, we formulate approximations that include higher-body interactions only among specific modes. For example, with $D = 4$, the decomposition including all one-body interactions and the three two-body interactions between the first and fourth, second and fourth, and third and fourth modes is defined as

$$\mathbf{T}_{i_1\dots i_4} \simeq \mathbf{P}_{i_1\dots i_4} = \frac{1}{Z_\chi} \circledast X^{(14)}_{i_1 i_4} \circledast X^{(24)}_{i_2 i_4} \circledast X^{(34)}_{i_3 i_4}, \tag{7}$$

where $X^{(d,4)}_{i_d i_4} = \mathrm{Exp}_\chi\left[E^{(d)}_{i_d} + E^{(d,4)}_{i_d,4} + E^{(4)}_4/3\right]$. We refer to the decomposition that reduces interactions between certain modes as the χ-deformed many-body approximation. Any model manifold $\mathscr{B}_\chi \subset \mathscr{H}_\chi$ of χ-deformed many-body approximation is represented as a linear condition on natural parameters $\boldsymbol{\theta}$. More specifically, it can be written as $\mathscr{B}_\chi = \{\mathbf{P} \in \mathscr{S}(D) \mid \theta^{\mathbf{P}}_{\boldsymbol{v}} = 0 \text{ if } \boldsymbol{v} \notin \Omega_B\}$ with appropriately given subindex set $\Omega_B \subset \Omega$. For example, in the (χ,n)-body approximation, the subset Ω_B is chosen such that $(\theta_{\boldsymbol{v}})_{\boldsymbol{v}\in\Omega_B}$ becomes the collection of all χ-deformed m-body parameters with $m \leq n$. Consequently, the χ-deformed many-body approximation $\mathbf{P} = \arg\min_{\mathbf{P}\in\mathscr{B}_\chi} D_\chi(\mathbf{T}, \mathbf{P})$ is an m-projection of $\mathbf{T}$ onto an e-flat manifold $\mathscr{B}_\chi$, which is a convex optimization problem and thus ensures a global minimum. Here, using the entropy $\varphi_\chi(\boldsymbol{\eta})$ derived in Section 4, the χ-divergence is given by

$$D_\chi(\mathbf{T}, \mathbf{P}) = \sum_{\boldsymbol{u}\in\Omega} \tilde{\chi}\left[\mathbf{T}_{\boldsymbol{u}}\right]\left\{\mathrm{Log}_\chi\left[\mathbf{T}_{\boldsymbol{u}}\right] - \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{u}}\right]\right\}, \tag{8}$$

where the χ-escort is defined as $\tilde{\chi}\left[\mathbf{T}_{\boldsymbol{u}}\right] = \chi[\mathbf{T}_{\boldsymbol{u}}]/\sum_{\boldsymbol{u}\in\Omega}\chi[\mathbf{T}_{\boldsymbol{u}}]$.

Connection to α-divergence When $\chi(s) = s^q$, the χ-divergence becomes the Tsallis divergence

$$D_q(\mathbf{T}, \mathbf{P}) = \frac{1}{1-q}\frac{1}{h_q(\mathbf{T})}\left[1 - \sum\nolimits_{\boldsymbol{u}\in\Omega} \mathbf{T}^q_{\boldsymbol{u}}\mathbf{P}^{1-q}_{\boldsymbol{u}}\right], \tag{9}$$

where $h_q(\mathbf{T}) = \sum_{\boldsymbol{u}\in\Omega}\mathbf{T}^q_{\boldsymbol{u}}$. It is straightforward to confirm its equivalence with Amari's α-divergence optimization, i.e., $\arg\min_{\mathbf{P}} D_q(\mathbf{T}, \mathbf{P}) = \arg\min_{\mathbf{P}} D^{\mathrm{Amari}}_q(\mathbf{T}, \mathbf{P})$ where the α-divergence is defined as $D^{\mathrm{Amari}}_q(\mathbf{T}, \mathbf{P}) = 1/(1-q)\{1 - \sum_{\boldsymbol{u}\in\Omega}\mathbf{T}^q_{\boldsymbol{u}}\mathbf{P}^{1-q}_{\boldsymbol{u}}\}$. The α-divergence-based decomposition is known for its robustness to outliers and mislabeled data and has been widely adopted in the signal processing and tensor communities (Cichocki et al., 2007; Kim et al., 2008; Hazan et al., 2007). In Appendix C, we empirically demonstrate that Tsallis-deformed many-body approximations also exhibit stability against additive noise. When $q \to 1$, the deformed product and the Tsallis divergence reduce to the ordinary product and the KL divergence, respectively. Consequently, the one-body approximation in Equation (5) becomes the ordinary rank-1 approximation that minimizes the KL divergence (Ghalamkari and Sugiyama, 2021). Further examples of the χ-divergence can be found in Appendix F.

3.2 Deformed low-rank approximation

The model capacity of the χ-deformed many-body approximation introduced in Section 3.1 can be increased by higher-order (larger-body) interactions. In machine learning, it is also common to enhance model capacity by latent variables, that is, random variables that are hidden via marginalization or summation over the variables. Recalling that in our framework, no distinction is made between tensor modes and random variables, we approximate a tensor $\mathbf{T} \in \mathscr{S}(D)$ by marginalization of a $(D+1)$-th order tensor $\mathbf{R} \in \mathbb{R}^{I_1\times\cdots\times I_D\times K}$, that is, $\mathbf{T}_{i_1\dots i_D} \simeq \mathbf{P}_{i_1\dots i_D} \coloneqq \sum_{k=1}^K \mathbf{R}_{i_1\dots i_D k}$. Here, $\mathbf{R}$ contains all one-body interactions and D two-body interactions between mode k and the other modes:

$$\mathbf{R}_{i_1\dots i_D k} = \frac{1}{Z_\chi} \circledast X^{(1)}_{i_1 k} \circledast X^{(2)}_{i_2 k} \circledast \dots \circledast X^{(D)}_{i_D k}, \tag{10}$$

for $X^{(d)}_{i_d k} = \mathrm{Exp}_\chi\left[E^{(d)}_{i_d} + E^{(d,D)}_{i_d,k} + E^{(D+1)}_k/D\right]$. We define the manifold $\mathscr{B}^{\mathrm{CP}}_\chi \subset \mathscr{S}(D+1)$ as the collection of tensors $\mathbf{R}$ that can be expressed in the form of Equation (10). This is a deformed extension of low-rank approximations since the tensor $\mathbf{P}$ becomes the rank-K approximation of $\mathbf{T}$ minimizing the KL-divergence and $X^{(d)}$ is called factor matrix when $\chi(s) = s$. Thus, we introduce the χ-deformed rank of the tensor $\mathbf{P}$ as $\mathrm{rank}_\chi(\mathbf{P}) = K$ defined by the smallest natural number K such that the tensor $\mathbf{P}$ can be written as a sum of χ-deformed products as above. We also refer to the tensor $\mathbf{P} = \arg\min_{\mathbf{P}\in\mathscr{M}^K_\chi} D_\chi(\mathbf{T},\mathbf{P})$ as a χ-deformed rank-K approximation of the tensor $\mathbf{T}$, where the set of deformed low-rank tensors is defined as $\mathscr{M}^K_\chi = \left\{\mathbf{P} \mid \mathrm{rank}_\chi(\mathbf{P}) \le K\right\}$. Please refer to Appendix G for further discussion on deformed Tucker and deformed Tensor Train structures.

Regularization via the Tsallis algebra As shown in Proposition 7, since the Tsallis-deformed product $x \circledast_q y$ reduces to $x + y - 1$ for $q \to 0$, the standard and the deformed rank of the deformed low-rank tensor are bounded by D and 1, respectively. It suggests that the parameter q in the Tsallis-deformed low-rank approximation controls the model capacity and provides implicit regularization, which is a property that does not appear in the Tsallis-deformed many-body approximation.

In contrast to the many-body approximation, the deformed low-rank approximation is a non-convex optimization problem since the latent variable k breaks convexity. However, when we define the function $\chi(\cdot)$ appropriately, the objective function of the deformed low-rank approximation can be relaxed into f-divergence, which is defined as $D^f(\mathbf{T},\mathbf{P}) = \sum_{\boldsymbol{u}} \mathbf{P}_{\boldsymbol{u}} f(\mathbf{T}_{\boldsymbol{u}}/\mathbf{P}_{\boldsymbol{u}})$ for a convex function f satisfying $f(1) = 0$. In such cases, we can obtain an approximate solution $\mathbf{P}$ through iterative convex subproblems, as introduced in Section 4.3.

4 OPTIMIZATION

We discuss the optimization procedures for the deformed many-body approximation and low-rank approximation. The former permits global optimization due to the flatness of the model manifold $\mathscr{B}_\chi$ while the latter requires non-convex optimization due to the non-flatness of $\mathscr{M}^K_\chi$.

4.1 Dual coordinate system on tensors

We perform the optimization using the coordinate transformation between $\boldsymbol{\theta} = (\theta_{\boldsymbol{v}})_{\boldsymbol{v}\in\Omega^+}$ and $\boldsymbol{\eta} = (\eta_{\boldsymbol{v}})_{\boldsymbol{v}\in\Omega^+}$ in $\mathscr{H}_\chi$. The transformation is carried out using the zeta function $\zeta : \Omega\times\Omega \to \{0,1\}$ and the Möbius function $\mu : \Omega\times\Omega \to \mathbb{Z}$, which are defined as follows:

$$\zeta(\boldsymbol{u},\boldsymbol{v}) = \begin{cases} 1 & \text{if } \boldsymbol{u} \le \boldsymbol{v}, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mu(\boldsymbol{u},\boldsymbol{v}) = \begin{cases} 1, & \text{if } \boldsymbol{u} = \boldsymbol{v}, \\ -\sum_{\boldsymbol{u}\le\boldsymbol{w}<\boldsymbol{v}} \mu(\boldsymbol{u},\boldsymbol{w}), & \text{if } \boldsymbol{u} < \boldsymbol{v}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{u} < \boldsymbol{v}$ indicates $\boldsymbol{u} \ne \boldsymbol{v}$ and $\boldsymbol{u} \le \boldsymbol{v}$. Using the Möbius inversion formula in Proposition 2, we obtain

$$\mathbf{P}_{\boldsymbol{i}} = \mathrm{Exp}_\chi\left[\sum\nolimits_{\boldsymbol{u}\in\Omega} \zeta(\boldsymbol{u},\boldsymbol{i})\theta_{\boldsymbol{u}}\right], \tag{11}$$

$$\theta_{\boldsymbol{u}} = \sum\nolimits_{\boldsymbol{w}\in\Omega} \mu(\boldsymbol{w},\boldsymbol{u})\,\mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{w}}\right], \tag{12}$$

where we define $\theta_\perp = -\psi_\chi(\boldsymbol{\theta})$. The Legendre transform of the χ-free energy $\psi_\chi(\boldsymbol{\theta})$ is given as $\varphi_\chi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}}\{\boldsymbol{\theta}^\top\boldsymbol{\eta} - \psi_\chi(\boldsymbol{\theta})\}$, which is referred to as the χ-entropy. These two convex functions, ψ_χ and φ_χ, provide dual representations $\boldsymbol{\eta} = \boldsymbol{\nabla}_{\boldsymbol{\theta}}\psi_\chi(\boldsymbol{\theta})$ and $\boldsymbol{\theta} = \boldsymbol{\nabla}_{\boldsymbol{\eta}}\varphi_\chi(\boldsymbol{\eta})$. Specifically, the χ-entropy $\varphi_\chi(\boldsymbol{\eta})$ is given as

$$\varphi_\chi(\boldsymbol{\eta}) = \sum\nolimits_{\boldsymbol{u}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{u}}\right]\mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{u}}\right], \tag{13}$$

where the χ-escort $\tilde{\chi}\left[\mathbf{T}_{\boldsymbol{v}}\right] = \chi[\mathbf{T}_{\boldsymbol{v}}]/\sum_{\boldsymbol{i}\in\Omega}\chi[\mathbf{T}_{\boldsymbol{i}}]$ is represented with $\boldsymbol{\eta}$-coordinates as follows:

$$\tilde{\chi}\left[\mathbf{P}_{\boldsymbol{v}}\right] = \sum_{\boldsymbol{w}\in\Omega} \mu(\boldsymbol{v},\boldsymbol{w})\eta_{\boldsymbol{w}}, \qquad \eta_{\boldsymbol{u}} = \sum_{\boldsymbol{w}\in\Omega} \zeta(\boldsymbol{u},\boldsymbol{w})\tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right]. \tag{14}$$

The χ-divergence in Equation (8) is the Bregman divergence generated by the convex function $\varphi_\chi(\boldsymbol{\eta})$. The derivations of the above results can be found in Propositions 4 to 6 in Appendix E. The representation in Equations (11) and (14) is a generalization of the transformation for the log-linear model on partially ordered set introduced by Sugiyama et al. (2018). The computational complexity of representing the tensor $\mathbf{P}$ in the θ- and η-coordinate systems scales linearly with the number of elements in the tensor $\mathbf{P}$. Please refer to Appendix F for an example of the transformation.

4.2 Global optimization for deformed many-body approximation

For a given tensor $\mathbf{T} \in \mathscr{S}(D)$, the subset of indices $\Omega_B \subset \Omega$, and the model manifold $\mathscr{B}_\chi = \{\mathbf{P} \in \mathscr{S}(D) \mid \theta^{\mathbf{P}}_{\boldsymbol{v}} = 0 \text{ if } \boldsymbol{v} \notin \Omega_B\} \subset \mathscr{H}_\chi$, the χ-deformed many-body approximation $\mathbf{P} = \arg\min_{\mathbf{P}\in\mathscr{B}_\chi} D_\chi(\mathbf{T},\mathbf{P})$ is a convex optimization problem as discussed in Section 3.1. Thus, we apply the Newton method based on the second-order derivatives of the objective function.

Table 1: Accuracy of the classification (top lines) and negative log-likelihood (bottom lines) for the first 4 datasets (in alphabetical order) out of a total of 25. The full results are available in Table 7. The bottom row reports the mean accuracy across all 25 datasets. The negative log-likelihood cannot be averaged because of its different scale. For proposed $q(\kappa)$-MBA, we report [optimal $q(\kappa)$].

Dataset	BM	KLTT	CNMF	EMCP	E^{2}MCPTTB	q-MBA	κ-MBA
AsiaL.	0.57 (0.00)	**0.59** (0.02)	0.57 (0.00)	0.57 (0.00)	0.50 (0.00)	0.57 [1.0]	0.57 [0.0]
	2.24 (0.00)	2.30 (0.06)	3.13 (0.09)	2.26 (0.00)	2.48 (0.16)	2.27 [1.0]	2.25 [0.7]
B.Scale	0.78 (0.00)	0.83 (0.00)	0.65 (0.07)	0.86 (0.01)	0.87 (0.00)	**0.89** [1.0]	**0.89** [0.0]
	7.32 (0.03)	7.48 (0.15)	7.42 (0.02)	8.98 (0.49)	7.10 (0.00)	7.04 [1.0]	**7.03** [0.3]
CarEval.	0.85 (0.04)	0.89 (0.00)	0.74 (0.01)	0.94 (0.02)	0.95 (0.02)	0.97 [1.0]	**0.98** [0.15]
	7.96 (0.15)	**7.75** (0.02)	8.19 (0.06)	7.78 (0.04)	7.85 (0.07)	7.79 [1.0]	7.82 [-0.7]
Chess2	0.16 (0.00)	0.28 (0.01)	0.24 (0.01)	0.44 (0.01)	0.57 (0.00)	**0.86** [0.95]	**0.86** [-0.2]
	13.13 (0.00)	12.56 (0.12)	12.77 (0.03)	11.95 (0.08)	11.64 (0.00)	11.42 [1.0]	**11.38** [-0.15]
Ave.Acc.	0.673	0.661	0.608	0.742	0.766	0.776	**0.779**

For $\boldsymbol{v}, \boldsymbol{w} \in \Omega_B$, the derivative of the χ-divergence is

$$\frac{\partial}{\partial \theta^{\mathbf{P}}_{\boldsymbol{v}}} D_\chi(\mathbf{T}, \mathbf{P}) = \eta^{\mathbf{P}}_{\boldsymbol{v}} - \eta^{\mathbf{T}}_{\boldsymbol{v}},$$
$$\frac{\partial}{\partial \theta^{\mathbf{P}}_{\boldsymbol{w}}} \frac{\partial}{\partial \theta^{\mathbf{P}}_{\boldsymbol{v}}} D_\chi(\mathbf{T}, \mathbf{P}) = \frac{\partial \eta^{\mathbf{P}}_{\boldsymbol{v}}}{\partial \theta^{\mathbf{P}}_{\boldsymbol{w}}} =: \mathbf{G}(\boldsymbol{\theta})_{\boldsymbol{vw}},$$

where Theorem 1 in Appendix E provides the χ-Fisher information matrix $\mathbf{G}(\boldsymbol{\theta}) \in \mathbb{R}^{|\Omega_B| \times |\Omega_B|}$ as:

$$\mathbf{G}(\boldsymbol{\theta})_{\boldsymbol{vw}} = \sum\nolimits_{\boldsymbol{u} \in \Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{u}}\right] \chi'\left[\mathbf{P}_{\boldsymbol{u}}\right] (\zeta(\boldsymbol{v}, \boldsymbol{u}) - \eta_{\boldsymbol{v}})(\zeta(\boldsymbol{w}, \boldsymbol{u}) - \eta_{\boldsymbol{w}}),$$

where the function χ' denotes the derivative of χ. Thus, the update rule is $\boldsymbol{\theta}^{\mathbf{P}}_B \leftarrow \boldsymbol{\theta}^{\mathbf{P}}_B - \mathbf{G}^{-1}(\boldsymbol{\eta}^{\mathbf{P}}_B - \boldsymbol{\eta}^{\mathbf{T}}_B)$ where $\boldsymbol{\theta}^{\mathbf{P}}_B = (\theta^{\mathbf{P}}_{\boldsymbol{v}})_{\boldsymbol{v} \in \Omega_B}$, $\boldsymbol{\eta}^{\mathbf{P}}_B = (\eta^{\mathbf{P}}_{\boldsymbol{v}})_{\boldsymbol{v} \in \Omega_B}$, and $\boldsymbol{\eta}^{\mathbf{T}}_B = (\eta^{\mathbf{T}}_{\boldsymbol{v}})_{\boldsymbol{v} \in \Omega_B}$. The free energy $\psi_\chi(\boldsymbol{\theta}^{\mathbf{P}})$ is computed numerically to ensure normalization, as described in Appendix G. Subsequently, we update the tensor $\mathbf{P}$ based on the updated $\boldsymbol{\theta}^{\mathbf{P}}$ using Equation (11), and the expectation parameter $\boldsymbol{\eta}^{\mathbf{P}}$ is then obtained from the updated $\mathbf{P}$ using Equation (14). Repeating this procedure finds the optimal tensor $\mathbf{P} \in \mathscr{B}_\chi$ that globally optimizes the χ-divergence. The entire algorithm is provided in Algorithm 1 in Appendix A, along with the L-BFGS method to remove the computational bottleneck of inverting $\mathbf{G}$ without loss of global optimality. In addition to the theoretical guarantee, we also numerically demonstrate the convergence in Appendix C.3.

4.3 *em*-deformed low-rank approximation

The set of deformed low-rank tensors $\mathscr{M}^K_\chi \subset \mathscr{H}_\chi$ is non-flat and thus the approximation $\arg\min_{\mathbf{P} \in \mathscr{M}^K_\chi} D_\chi(\mathbf{T}, \mathbf{P})$ for given tensor $\mathbf{T}$ requires non-convex optimization. However, if there exists a convex function f such that $\arg\min_{\mathbf{P}} D^f(\mathbf{T}, \mathbf{P}) = \arg\min_{\mathbf{P}} D_\chi(\mathbf{T}, \mathbf{P})$, where $D^f(\mathbf{T}, \mathbf{P})$ is a f-divergence introduced in Section 3.2, we can approximately obtain the deformed low-rank approximation $\mathbf{P}$ by iterative convex optimization with convergence guarantee. One example where this relation holds is when $\chi(s) = s^q$ and $f(s) = (s^q - qs - 1 + q)/(q^2 - q)$. In this case, the corresponding χ-divergence and f-divergence are the Tsallis divergence and Amari's α-divergence, respectively. It remains an open question to identify χ-divergences other than the Tsallis divergence for which optimization is equivalent to that of an f-divergence.

We relax the projection from $\mathbf{T} \in \mathscr{S}(D)$ to $\mathscr{M}^K_\chi \subset \mathscr{S}(D)$, which requires non-convex optimization, into an iterative convex optimization problem in higher order tensor space $\mathscr{S}(D+1)$. Specifically, we consider the data manifold $\mathscr{D} = \left\{\mathbf{Q} \mid \sum_k \mathbf{Q}_{i_1 \dots i_D k} = \mathbf{T}_{\boldsymbol{i}}\right\} \subset \mathscr{S}(D+1)$, consisting of tensors whose marginalization yields the tensor $\mathbf{T}$, and the e-flat manifold $\mathscr{B}^{\mathrm{CP}}_\chi \subset \mathscr{S}(D+1)$ consisting of tensors in the form of Equation (10).

The m-flatness of the manifold $\mathscr{D}$ is discussed in Proposition 8 in Appendix E. We alternately perform e-projection from a point on $\mathscr{B}^{\mathrm{CP}}_\chi$ to $\mathscr{D}$ and m-projection from a point on $\mathscr{D}$ to $\mathscr{B}^{\mathrm{CP}}_\chi$. The former is called e-step and the latter is called m-step. For tensors $\mathbf{R} \in \mathscr{B}^{\mathrm{CP}}_\chi$, these steps monotonically decrease the objective function by minimizing the upper bound $D^f(\mathbf{Q}, \mathbf{R}) \geq D^f(\mathbf{T}, \mathbf{P})$, as derived in Proposition 9 in Appendix E. This inequality does not hold for general χ-divergences, making the extension of the algorithm to general χ-divergences nontrivial. Each e- and m-step is explicitly described as follows.

e-step Fixing tensor $\mathbf{R}$, we minimize the upper bound $D^f(\mathbf{Q}, \mathbf{R})$ for $\mathbf{Q} \in \mathscr{D}$. This is an e-projection

onto the m-flat manifold $\mathscr{D}$ in the coordinate system of the standard exponential family $\mathscr{H}$, thus, convex optimization. As seen in Proposition 10 in Appendix E, this optimization has a closed-form solution given by

$$\mathbf{Q}^*_{i_1 \dots i_D k} = \frac{\mathbf{T}_{\boldsymbol{i}} \mathbf{R}_{i_1 \dots i_D k}}{\sum_k \mathbf{R}_{i_1 \dots i_D k}}. \tag{15}$$

m-step Fixing tensor $\mathbf{Q}$, we minimize $D^f(\mathbf{Q}, \mathbf{R})$ for $\mathbf{R} \in \mathscr{B}^{\mathrm{CP}}_{\chi}$. This is a deformed many-body approximation, thus, a convex optimization problem.

The entire procedure is illustrated in Algorithm 2 in Appendix A, along with the computational complexity. Although the deformed extension of the EM algorithm was introduced by Qin and Priebe (2013), our above information-geometric analysis naturally guarantees the convexity of each step.

As shown in Theorem 3 in Appendix E, convergence is guaranteed by the fact that the e-step tightens the bound, i.e., $D^f(\mathbf{T}, \mathbf{P}) = D^f(\mathbf{Q}^*, \mathbf{R})$. The EM-based ordinary CP decomposition (Huang and Sidiropoulos, 2017) is recovered when $\chi(s) = s$. We emphasize that if we use the ordinary outer product instead of the deformed product, the m-step remains a non-convex optimization problem.

4.4 Computational complexity

While the second-order approach converges in fewer iterations, computing the inverse of $\mathbf{G}$ becomes the main computational bottleneck, requiring $|\Omega_B|^3$ operations. This difficulty can be eliminated by employing the L-BFGS method (Liu and Nocedal, 1989), which reduces the complexity of computing $\mathbf{G}^{-1}$ to $O(\mu|\Omega_B|)$, where μ denotes the number of past updates stored in memory. When the L-BFGS method is applied, the bottleneck is computing all expectation parameters $\boldsymbol{\eta}$ each iteration using Equation (14), which requires access to all elements of the tensor. The total computational complexity, including that of the existing sparse q-CP decomposition optimizing the α-divergence (Ghalamkari et al., 2024), is summarized in Table 2. It is an interesting future direction to reduce the computational complexity of updating the expectation parameters by employing appropriate sampling methods, enabling the proposed method to be applied in sparse settings.

Models	Complexity
MBA (Newton)	$O(\max(\|\Omega_B\|^3, \|\Omega\|))$
MBA (LBFGS)	$O(\|\Omega\|)$
q-DCP (LBFGS)	$O(KI^{D-1})$
q-CP	$O(DNK^2)$

Table 2: Computational complexity of each iteration for the tensor dimension D, number of non-zero values N, number of parameters $|\Omega_B|$, number of elements of the tensor $|\Omega|$, and rank K assuming the tensor size $I \times I \times \cdots \times I$.

5 NUMERICAL EVALUATION

We apply the proposed methods to PMF estimation, which is a well-established application of tensor factorization (Kargas et al., 2018), to evaluate their effectiveness. Specifically, given discrete samples $\mathcal{X} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^N)$ for $\boldsymbol{x}^n \in \Omega = [I_1] \times \cdots \times [I_D]$, we construct the normalized tensor $\mathbf{T} \in \mathbb{R}^{I_1 \times \cdots \times I_D}_{\geq 0}$ such that $\mathbf{T}_{\boldsymbol{i}} = 1/N \sum_{n=1}^N \delta(\boldsymbol{x}^n, \boldsymbol{i})$ where $\delta(\boldsymbol{x}^n, \boldsymbol{i})$ is 1 if $\boldsymbol{x}^n = \boldsymbol{i}$, 0 otherwise. The reconstructed tensor $\mathbf{P}$, obtained by decomposing $\mathbf{T}$, approximates the underlying distribution of the data, with each entry $\mathbf{P}_{\boldsymbol{i}}$ representing the probability $p(x_1 = i_1, \dots, x_D = i_D)$. While we mainly focus on the CarEvaluation dataset here due to space constraints, additional results on diverse datasets, along with detailed experimental setups, can be found in Appendices B and C.

To investigate the optimization performance of the proposed methods, we plot the training error of each iteration of the Tsallis-deformed many-body approximation (q-MBA) and Tsallis-deformed low-rank approximation (q-DCP) in Figure 2(**a**). As a benchmark, we also show the corresponding curves obtained by the double-bounded EM-based CP decomposition optimizing the α-divergence (q-CP) (Ghalamkari et al., 2024). For a fair comparison, the (deformed) ranks are chosen such that the number of parameters in the (deformed) low-rank model matches that of the many-body approximation. We plotted the results from 10 random initializations for both q-DCP and q-CP, and from a single run for q-MBA. The results indicate that q-MBA effectively reduces the α-divergence, whereas q-DCP with smaller q fails to achieve lower training error than the baseline, since smaller q provides a stronger regularization effect, as discussed in Section 3.2.

To assess the effect of robustness and regularization induced by the deformed formulation on generalization performance, we evaluate the quality of the obtained mass via classification accuracy and negative log-likelihood on the validation data, as shown in Figure 2(**b**), varying the (deformed) rank K and the parameter q. For classification, we obtained class-conditional probabilities from the estimated mass and assigned each sample to the class with the highest conditional probability. The results show that the q-DCP avoids overfitting while q-CP tends to overfit easily. The better generalization performance of q-DCP is also observed in the noisy COIL image reconstruction task, as summarized in Appendix C.

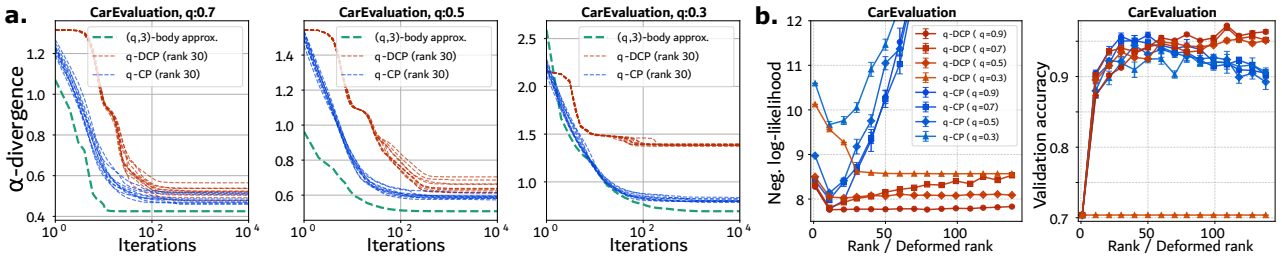


Figure 2: The training error per iteration for different q values (**a**), and classification accuracy and negative log-likelihood on the validation data for varying (deformed) ranks and q values (**b**). Results for additional six datasets can be found in Appendix C.

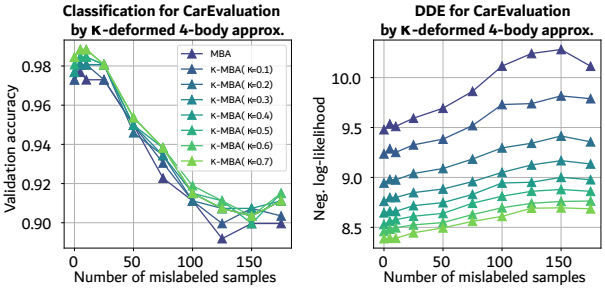


Figure 3: The performance of κ-MBA with mislabeled samples in the training dataset. Results for additional six datasets can be found in Appendix C.

To evaluate the practical usefulness of the q-MBA and Kaniadakis-deformed many-body approximations (κ-MBA), we use them for PMF estimation considering 25 datasets, and evaluate the estimated mass using negative log-likelihood and classification performance. We restrict the datasets to those whose associated tensors have fewer than ten million entries. Due to space limitations, Table 1 reports the results on the test data for the first four datasets in alphabetical order, including results from the following tensor-based baseline methods for PMF estimation, Born machine (BM), KL-divergence-based Tensor Train (KLTT) (Glasser et al., 2019), pairwise tensor method (CNMF) (Ibrahim and Fu, 2021), EM-based CP decomposition (Yeredor and Haardt, 2019), and mixture tensor-based method (E^2MCPTTB) (Ghalamkari et al., 2024). The parameters q, κ, and body in our method, as well as the ranks in the baseline methods, are tuned to maximize the validation score. Notably, both q-MBA and κ-MBA are guaranteed to achieve global optimality and are free from initialization dependency, while attaining scores comparable to the baseline methods.

Interestingly, the optimal κ values of κ-MBA that minimize the validation log-likelihood are often greater than zero, which suggests that minimizing the Kaniadakis-divergence on the training data can improve the test likelihood more effectively than directly maximizing the training log-likelihood. To further investigate the property of the κ-MBA, we randomly flipped the class labels of the training samples and evaluated the classification performance and log-likelihood on the test dataset in Figure 3. Although label noise degrades the quality of the estimated mass, we observe that the increase in negative log-likelihood becomes less steep for larger κ values. As seen in Appendix C, this phenomenon is consistently reproduced for the κ-MBA across many datasets, whereas it does not appear for the q-MBA.

6 CONCLUSION

We reformulated tensor decomposition using deformed algebra and deformed information geometry, establishing that the extended many-body approximation can be globally optimized under various divergences. Furthermore, we developed an *em*-procedure for the Tsallis-deformed extension of non-convex low-rank approximation as iterative convex subproblems. Finally, we discovered that the Tsallis-deformed algebra induces regularization in deformed low-rank models, preventing overfitting, while the deformed many-body approximation retains its capacity regardless of the value of q.

Further applications Since one-body approximations correspond to mean-field approximations, our deformed formulation may inspire Variational Inference (Zhang and Yang, 2024), Variational Gaussian Processes (Hamelijnck et al., 2021), and Boltzmann Machines (Kappen and de Borja Rodríguez Ortiz, 1997) by providing a χ-divergence-based robust mean-field foundation. As low-rank tensor methods connect to probabilistic circuits (Loconte et al., 2025) and graphical models (Robeva and Seigal, 2018), our generalized formulation naturally yields their deformed extension, potentially benefiting from the Tsallis-based implicit regularization, which is expected to aid generalization, as suggested by our empirical results in Figure 2(**b**).

Limitations Our formulation requires accessing all tensor elements in each iteration to update $\boldsymbol{\eta}$, which prevents the method from scaling to large datasets.

Acknowledgements

This work was supported by the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation, Grant Number NNF21SA0069429 (GK). MM and PT were supported by the Novo Nordisk Foundation, Grant Number NNF23OC0083524.

References

Acar, E., Dunlavy, D. M., Kolda, T. G., and Mørup, M. (2011). Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56.

Alefeld, G. E., Potra, F. A., and Shi, Y. (1995). Algorithm 748: enclosing zeros of continuous functions. *ACM Trans. Math. Softw.*, 21(3):327–344.

Amari, S.-i. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276.

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

Amari, S.-i. and Ohara, A. (2011). Geometry of q-exponential family of probability distributions. *Entropy*, 13(6):1170–1185.

Amari, S.-i., Ohara, A., and Matsuzoe, H. (2012). Geometry of deformed exponential families: Invariant, dually-flat and conformal geometries. *Physica A: Statistical Mechanics and its Applications*, 391(18):4308–4319.

Amid, E., Warmuth, M. K., and Srinivasan, S. (2019). Two-temperature logistic regression based on the tsallis divergence. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2388–2396. PMLR.

Anteneodo, C. and Plastino, A. R. (1999). Maximum entropy approach to stretched exponential probability distributions. *Journal of Physics A: Mathematical and General*, 32(7):1089.

Armijo, L. (1966). Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3.

Bienaymé, I.-J. (1853). *Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés*. Imprimerie de Mallet-Bachelier.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Borges, E. P. (2004). A possible deformed algebra and calculus inspired in nonextensive thermostatistics. *Physica A: Statistical Mechanics and its Applications*, 340(1):95–101. News and Expectations in Thermostatistics.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217.

Chanu, T. R., Singh, T. R., and Singh, K. M. (2023). A new algorithm for removing salt and pepper noise from color medical images. *Multimedia Tools and Applications*, 82(16):24991–25013.

Chege, J. K., Grasis, M. J., Manina, A., Yeredor, A., and Haardt, M. (2022). Efficient probability mass function estimation from partially observed data. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pages 256–262.

Chege, J. K., Grasis, M. J., Yeredor, A., and Haardt, M. (2023). Bayesian estimation of a probability mass function tensor with automatic rank detection. In *2023 IEEE 9th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 211–215.

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

Chung, J., Kannappan, P., Ng, C. T., and Sahoo, P. (1989). Measures of distance between probability distributions. *Journal of Mathematical Analysis and Applications*, 138(1):280–292.

Cichocki, A., Cruces, S., and Amari, S.-i. (2011). Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170.

Cichocki, A., Zdunek, R., and Amari, S.-i. (2006). Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *International conference on independent component analysis and signal separation*, pages 32–39. Springer.

Cichocki, A., Zdunek, R., Choi, S., Plemmons, R., and Amari, S.-i. (2007). Non-negative tensor factorization using alpha and beta divergences. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 3, pages III–1393–III–1396.

Clementi, F., Gallegati, M., and Kaniadakis, G. (2007). κ-generalized statistics in personal income distribution. *The European Physical Journal B*, 57(2):187–193.

da Silva, J. A., Cirillo, M. A., and Manuel, L. (2024). Maximum L_q-likelihood estimation: A study of en-

tropy behavior for the pareto-exponential distribution with application. *Journal of Statistical Theory and Practice*, 18(3):44.

Flores, P., Chege, J. K., Usevich, K., Haardt, M., Yeredor, A., and Brie, D. (2023). Probability mass function estimation approaches with application to flow cytometry data analysis. In *2023 IEEE 9th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 451–455.

Fujimoto, Y. and Murata, N. (2012). Nonnegative matrix factorization via generalized product rule and its application for classification. In Theis, F., Cichocki, A., Yeredor, A., and Zibulevsky, M., editors, *Latent Variable Analysis and Signal Separation*, pages 263–271, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ghalamkari, K., Hinrich, J. L., and Mørup, M. (2024). E^2M: Double bounded α-divergence optimization for tensor-based discrete density estimation. *arXiv preprint arXiv:2405.18220*.

Ghalamkari, K., Hinrich, J. L., and Mørup, M. (2025). Non-negative tensor low-rank decompositions through the lens of information geometry. In *Proceedings of the AAAI-25 Workshop on CoLoRAI: Connecting Low-Rank Representations in AI*.

Ghalamkari, K. and Sugiyama, M. (2021). Fast tucker rank reduction for non-negative tensors using mean-field approximation. In *Advances in Neural Information Processing Systems*, volume 34, pages 443–454, Virtual Event.

Ghalamkari, K. and Sugiyama, M. (2022). Fast rank-1 NMF for missing data with KL divergence. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 2927–2940, Virtual Event.

Ghalamkari, K. and Sugiyama, M. (2023). Non-negative low-rank approximations for multi-dimensional arrays on statistical manifold. *Information Geometry*, 6:257–292.

Ghalamkari, K., Sugiyama, M., and Kawahara, Y. (2023). Many-body approximation for non-negative tensors. In *Advances in Neural Information Processing Systems*, volume 36, pages 74077–74102, New Orleans, US.

Gill, P. E., Murray, W., and Wright, M. H. (2019). *Practical Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Giméenez, P., Guarracino, L., and Galea, M. (2022). Maximum L_q-likelihood estimation in functional measurement error models. *Statistica Sinica*, 32(3):1723–1743.

Glasser, I., Sweke, R., Pancotti, N., Eisert, J., and Cirac, I. (2019). Expressive power of tensor-network factorizations for probabilistic modeling. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix computations*. JHU press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Graham, R. L., Knuth, D. E., and Patashnik, O. (1994). *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Longman Publishing Co., Inc., USA, 2nd edition.

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

Hamelijnck, O., Wilkinson, W., Loppi, N., Solin, A., and Damoulas, T. (2021). Spatio-temporal variational gaussian processes. *Advances in Neural Information Processing Systems*, 34:23621–23633.

Hanel, R. and Thurner, S. (2011). A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions. *Europhysics Letters*, 93(2):20006.

Hansen, S., Plantenga, T., and Kolda, T. G. (2015). Newton-based optimization for Kullback–Leibler nonnegative tensor factorizations. *Optimization Methods and Software*, 30(5):1002–1029.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hazan, T., Hardoon, R., and Shashua, A. (2007). pLSA for sparse arrays with Tsallis pseudo-additive divergence: Noise robustness and algorithm. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8.

Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernandez-Lobato, D., and Turner, R. (2016). Black-box α- divergence minimization. In Balcan, M. F.

and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1511–1520, New York, New York, USA. PMLR.

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

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

Hitchcock, F. L. (1927). The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189.

Hitomi, M. and Ohzeki, M. (2025). Typical reconstruction limit and phase transition of maximum entropy method. *Physical Review Research*, 7(4):043091.

Hollender, A., Lawrence, C., and Segal-Halevi, E. (2025). Computing approximate roots of monotone functions. *Journal of Complexity*, 88:101930.

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.

Ibrahim, S. and Fu, X. (2021). Recovering joint probability of discrete random variables from pairwise marginals. *IEEE Transactions on Signal Processing*, 69:4116–4131.

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

Kaniadakis, G. (2013). Theoretical foundations and mathematical formalism of the power-law tailed statistical distributions. *Entropy*, 15(10):3983–4010.

Kaniadakis, G., Baldi, M. M., Deisboeck, T. S., Grisolia, G., Hristopulos, D. T., Scarfone, A. M., Sparavigna, A., Wada, T., and Lucia, U. (2020). The κ-statistics approach to epidemiology. *Scientific reports*, 10(1):19949.

Kaniadakis, G., Lissia, M., and Scarfone, A. (2005). Two-parameter deformations of logarithm, exponential, and entropy: A consistent framework for generalized statistical mechanics. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 71(4):046128.

Kappen, H. and de Borja Rodríguez Ortiz, F. (1997). Boltzmann machine learning using mean field theory and linear response correction. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press.

Kargas, N., Sidiropoulos, N. D., and Fu, X. (2018). Tensors, learning, and "Kolmogorov extension" for finite-alphabet random vectors. *IEEE Transactions on Signal Processing*, 66(18):4854–4868.

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.

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

Kossaifi, J., Lipton, Z. C., Kolbeinsson, A., Khanna, A., Furlanello, T., and Anandkumar, A. (2020). Tensor regression networks. *Journal of Machine Learning Research*, 21(123):1–21.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

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

Legendre, A.-M. (1789). *Mémoire sur la manière de distinguer les maxima des minima dans le calcul des variations*, volume 1787. Imprimerie royale, Paris.

Li, C. T. and Farnia, F. (2023). Mode-seeking divergences: theory and applications to gans. In *International Conference on Artificial Intelligence and Statistics*, pages 8321–8350. PMLR.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528.

Loconte, L., Mari, A., Gala, G., Peharz, R., de Campos, C., Quaeghebeur, E., Vessio, G., and Vergari, A. (2025). What is the relationship between tensor factorizations and circuits (and how can we exploit it)? *Transactions on Machine Learning Research*.

Masrani, V., Brekelmans, R., Bui, T., Nielsen, F., Galstyan, A., Ver Steeg, G., and Wood, F. (2021). q-paths: Generalizing the geometric annealing path using power means. In de Campos, C. and Maathuis, M. H., editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1938–1947. PMLR.

Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, page 362–369, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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

Murg, V., Verstraete, F., Schneider, R., Nagy, P. R., and Legeza, O. (2015). Tree tensor network state with variable tensor order: An efficient multireference method for strongly correlated systems. *Journal of Chemical Theory and Computation*, 11(3):1027–1036.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Möbius, A. (1832). Über eine besondere art von umkehrung der reihen. *Journal für die reine und angewandte Mathematik*, 9:105–123.

Naudts, J. (2002). Deformed exponentials and logarithms in generalized thermostatistics. *Physica A: Statistical Mechanics and its Applications*, 316(1-4):323–334.

Naudts, J. (2011). *Generalised Thermostatistics*. Springer Science & Business Media.

Nayar and Murase, H. (1996). Columbia object image library: COIL-100. Technical Report CUCS-006-96, Department of Computer Science, Columbia University.

Novikov, G. S., Panov, M. E., and Oseledets, I. V. (2021). Tensor-train density estimation. In de Campos, C. and Maathuis, M. H., editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1321–1331. PMLR.

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., and Moore, J. H. (2017). Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, 10(36):1–13.

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

Pistone, G. and Shoaib, M. (2023). Kaniadakis's information geometry of compositional data. *Entropy*, 25(7):1107.

Qin, Y. and Priebe, C. E. (2013). Maximum L_q-likelihood estimation via the expectation-maximization algorithm: A robust estimation of mixture models. *Journal of the American Statistical Association*, 108(503):914–928.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 547–562. University of California Press.

Robeva, E. and Seigal, A. (2018). Duality of graphical models and tensor networks. *Information and Inference: A Journal of the IMA*, 8(2):273–288.

Romano, J. D., Le, T. T., La Cava, W., Gregg, J. T., Goldberg, D. J., Chakraborty, P., Ray, N. L., Himmelstein, D., Fu, W., and Moore, J. H. (2021). Pmlb v1.0: an open source dataset collection for benchmarking machine learning methods. *arXiv preprint arXiv:2012.00058v2*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Sra, S. and Dhillon, I. (2005). Generalized nonnegative matrix approximations with Bregman divergences. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Sugiyama, M., Nakahara, H., and Tsuda, K. (2017). Tensor balancing on statistical manifold. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 3270–3279, Sydney, Australia.

Sugiyama, M., Nakahara, H., and Tsuda, K. (2018). Legendre decomposition for tensors. In *Advances in Neural Information Processing Systems 31*, pages 8825–8835, Montréal, Canada.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52(1):479–487.

Tsallis, C. (1999). Nonextensive statistics: theoretical, experimental and computational evidences and connections. *Brazilian Journal of Physics*, 29:1–35.

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

Van Erven, T. and Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wada, T. and Scarfone, A. M. (2023). On the Kaniadakis distributions applied in statistical physics and natural sciences. *Entropy*, 25(2).

Wu, Z.-C., Huang, T.-Z., Deng, L.-J., Dou, H.-X., and Meng, D. (2022). Tensor wheel decomposition and its tensor completion application. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27008–27020. Curran Associates, Inc.

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

Zhang, M., Bird, T., Habib, R., Xu, T., and Barber, D. (2019). Variational f-divergence minimization. *arXiv preprint arXiv:1907.11891*.

Zhang, Y. and Yang, Y. (2024). Bayesian model selection via mean-field variational approximation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):742–770.

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

Zheng, Y.-B., Huang, T.-Z., Zhao, X.-L., Zhao, Q., and Jiang, T.-X. (2021). Fully-connected tensor network decomposition and its application to higher-order tensor completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11071–11078.

Zhu, L., Shah, H., Wang, H., Nagai, Y., and White, M. (2025). q-exponential family for policy optimization. In *The Thirteenth International Conference on Learning Representations*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes. All mathematical statements and proposed algorithm are formally described in Appendices A , D, and E with clear settings.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes. The computational complexity and running time per iteration are stated in Appendices A and C]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes. We provide the README file to describe them.]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, the statements, along with their assumptions, are formally given in Appendix E.]
   (b) Complete proofs of all theoretical results. [Yes. Complete proofs of theoretical results can be found in Appendix E]
   (c) Clear explanations of any assumptions. [Yes. The assumption of non-negativity and normalization is stated in Section 3.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes. Please see Appendix B]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes. The information is also available in Appendix B]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes. We detailed them in Appendix C.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes. The information is available in Appendix B]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes. We cite the baseline methods and dataset URL.]
   (b) The license information of the assets, if applicable. [Yes. It is available in Table 5]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes. The code can be found in the supplemental material.]
   (d) Information about consent from data providers/curators. [Not Applicable. We used only open datasets.]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable.]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable.]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable.]

Supplementary Materials for Deformed Decomposition for Non-negative Tensors

Table of Contents

Table 3: List of notations

Symbol	Description
$\mathbf{T}$, $\mathbf{P}$	Given tensor (data), approximated tensor (model)
$\mathbf{P}^{\leq m}$	m-body tensor, i.e., $\mathbf{P}^{\leq m} \in \mathscr{B}_\chi^{\leq m}$
D	Tensor order (number of modes)
I_d	Dimension of mode d
$[n]$	The set of natural number smaller than n, i.e., $[n] = \{1, 2, \ldots, n\}$
$\Omega = [I_1] \times \cdots \times [I_D]$	Index set of tensor entries; sample space
$\boldsymbol{i} = (i_1, \ldots, i_D)$	Tensor index; $\boldsymbol{i} \in \Omega$
$\boldsymbol{\theta}$, $\boldsymbol{\eta}$	Natural and expectation parameters of the χ-exponential family
$\psi_\chi(\boldsymbol{\theta})$	χ-free energy (log-partition function)
$\varphi_\chi(\boldsymbol{\eta})$	χ-entropy (Legendre dual of ψ_χ)
$\mathrm{Exp}_\chi[\cdot]$, $\mathrm{Log}_\chi[\cdot]$	χ-exponential and χ-logarithmic functions (Equation (1))
$\chi(s)$	Positive deformation function (e.g., s^q, $s/\cosh(\kappa \log s)$)
χ'	Derivative of χ
$\tilde{\chi}[\mathbf{P}]$	χ-escort distribution: $\tilde{\chi}[\mathbf{P}] = \chi[\mathbf{P}] / \sum_{\boldsymbol{i} \in \Omega} \chi[\mathbf{P}_{\boldsymbol{i}}]$
$\circledast_\chi$	χ-deformed product satisfying $\mathrm{Exp}_\chi[a+b] = \mathrm{Exp}_\chi[a] \circledast_\chi \mathrm{Exp}_\chi[b]$
$\otimes$	Standard outer product
$E_{\boldsymbol{i}}$	Energy function of tensor element $\boldsymbol{i} \in \Omega$ (Equation (3))
$\mathscr{B}_\chi$	Model manifold of χ-deformed many-body approximation
$\mathscr{B}_\chi^{\leq m}$	Model manifold of (χ, m)-body approximation
$\mathscr{M}_\chi^{\leq K}$	Set of χ-deformed tensors with rank at most K
$\mathrm{rank}_\chi(\mathbf{P})$	χ-deformed tensor rank of $\mathbf{P}$ (Definition 3)
$\mathrm{rank}_{\mathrm{CP}}(\mathbf{P})$	Canonical (CP) tensor rank of $\mathbf{P}$ (Definition 2)
$D_\chi(\mathbf{T}, \mathbf{P})$	χ-divergence between tensors (Equation (8))
$D_\chi^*(\mathbf{T}, \mathbf{P})$	Dual χ-divergence, i.e., $D_\chi^*(\mathbf{T}, \mathbf{P}) = D_\chi(\mathbf{P}, \mathbf{T})$
D_q, D_κ	Tsallis (q) and Kaniadakis (κ) divergences (Equations (37) and (38))
D^f	f-divergence (Definition 5)
m-, e-projection	Projection minimizing $D_\chi(p, q)$ or $D_\chi(q, p)$, respectively, for given p
$\mathbf{G}$	χ-Fisher information matrix (Definition 11)
Z_χ	Normalization constant, i.e., $Z_\chi^{-1} = \mathrm{Exp}_\chi[-\psi_\chi(\boldsymbol{\theta})]$ (deformed partition function)
q, κ	Tsallis and Kaniadakis deformation parameters
$\perp = (1, \ldots, 1)$	Least element in the sample space Ω
$\mathscr{H}$	Standard exponential family
$\mathscr{H}_\chi$	χ-exponential family
$\mathscr{D}$	Data manifold: $\left\{\mathbf{Q} \mid \sum_k \mathbf{Q}_{i_1, \ldots, i_D, k} = \mathbf{T}_{i_1, \ldots, i_D}\right\}$ for given $\mathbf{T}$
$\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_{\geq 0}$	Sets of natural numbers, real numbers, and non-negative real numbers
$\mu(\cdot, \cdot)$, $\zeta(\cdot, \cdot)$	Möbius and zeta functions (Definition 13)
$\delta(\cdot, \cdot)$	Kronecker delta: $\delta(x, y) = 1$ if $x = y$, 0 otherwise
$\mathscr{S}(D)$	Set of non-negative normalized tensors of size $I_1 \times \cdots \times I_D$
$\mathscr{S}(D{+}1)$	Set of non-negative normalized tensors of size $I_1 \times \cdots \times I_D \times K$
$\lVert \cdot \rVert_F$	Frobenius norm

Algorithm 1: χ-deformed many-body approximation

DeformedManyBodyApproximation($\mathbf{T}$, Ω_B, $\boldsymbol{\theta}^{t=0}$) // $\boldsymbol{\theta}^{t=0}$ is the initial value of $\boldsymbol{\theta}^{\mathbf{P}}$
  Compute $\eta^{\mathbf{T}}_{\boldsymbol{v}}$ for $\boldsymbol{v} \in \Omega_B$ using Equation (14).
  repeat
    Compute $\psi_\chi(\boldsymbol{\theta}^{\mathbf{P}})$ by numerically solving Equation (41) using bisection method;
    Compute $\mathbf{P}$ using the current parameter $\boldsymbol{\theta}^{\mathbf{P}}$ and $\psi_\chi(\boldsymbol{\theta}^{\mathbf{P}})$ with Equation (2);
    Compute $\eta^{\mathbf{P}}_{\boldsymbol{v}}$ for $\boldsymbol{v} \in \Omega_B$ from $\mathbf{P}$ using Equation (14);
    Compute the χ-Fisher information matrix $\mathbf{G}(\boldsymbol{\theta}^{\mathbf{P}})$ using Equation (25);
    Estimate appropriate learning-rate β by line search;
    $\boldsymbol{\theta}^{\mathbf{P}}_B \leftarrow \boldsymbol{\theta}^{\mathbf{P}}_B - \beta \mathbf{G}^{-1}(\boldsymbol{\eta}^{\mathbf{P}}_B - \boldsymbol{\eta}^{\mathbf{T}}_B)$;
  until *Convergence*;
  return $\mathbf{P}$, $\boldsymbol{\theta}^{\mathbf{P}}$

Algorithm 2: *em*-based χ-deformed rank-K approximation

DeformedLowrankApproximation($\mathbf{T}$, K)
  Initialize $\mathbf{R}$ and $\boldsymbol{\theta}$; // e.g. random normalized tensor $\mathbf{R}$ and $\boldsymbol{\theta} = \mathbf{0}$
  Define Ω_B appropriately for the desired low-rank structure;
  repeat
    Compute $\mathbf{Q}$ from $\mathbf{R}$ using Equation (15); // e-step
    $\mathbf{R}$, $\boldsymbol{\theta} \leftarrow$ DeformedManyBodyApproximation ($\mathbf{Q}$, Ω_B, $\boldsymbol{\theta}$) ; // m-step
  until *Convergence*;
  return $\mathbf{R}$

A ALGORITHM AND IMPLEMENTATION DETAILS

We present the proposed χ-deformed many-body approximation and *em*-based χ-deformed low-rank approximation as pseudocode in Algorithms 1 and 2, respectively. The proposed *em*-algorithm is applicable only when the objective function can be relaxed to an f-divergence, as discussed in Section 4.3. Although the algorithms are based on second-order derivatives, they can also be implemented with only first-order derivatives by replacing the χ-Fisher information matrix $\mathbf{G}$ with the identity matrix. We also note that the second derivative, namely the χ-Fisher information matrix $\mathbf{G}(\boldsymbol{\theta}) \in \mathbb{R}^{|\Omega_B| \times |\Omega_B|}$, defines the metric in the model space $\mathscr{B}_\chi$. Optimization with the metric is commonly referred to as the natural gradient method (Amari, 1998).

In Algorithm 1, we need to obtain the free energy ψ_χ by numerically solving Equation (41) to ensure normalization. The optimal unique value of ψ_χ can be obtained efficiently by the bisection method as detailed in Section G. Specifically, we used the accelerated bisection method, `toms748` (Alefeld et al., 1995), in the SciPy library (Virtanen et al., 2020). The learning rate $\beta > 0$ is determined at each iteration by use of line search (Armijo, 1966). We used the `scipy.optimize.line_search` function with default parameters.

To ensure numerical stability of the updates, we apply two standard techniques. First, we stabilize the update by adding a damping term to the χ-Fisher information matrix. Specifically, we estimate the maximum eigenvalue $\lambda_{\max}$ of $\mathbf{G}$ using the power method, and add $\max(10^{-12}, 10^{-6} \cdot \lambda_{\max})$ to the diagonal elements of $\mathbf{G}$. This procedure, known as damping, is a widely used heuristic in Newton methods (Gill et al., 2019). Second, when computing the pseudo-inverse of $\mathbf{G}$, we truncate small singular values below a threshold $\mathrm{rcond} = 10^{-8}$ to remove numerically insignificant directions and further improve stability (Golub and Van Loan, 2013).

These algorithms were regarded as converged when the difference between the current and previous cost function values was less than 10^{-8}. For Algorithm 1, the initial value of $\boldsymbol{\theta}$ was set to the natural parameter of the uniform tensor, i.e., $(\boldsymbol{\theta}^{t=0}_{\boldsymbol{v}})_{\boldsymbol{v} \in \Omega_B} = 0$. In Algorithm 2, the initial value for each m-step iteration is set to the natural parameter obtained from the previous m-step. In the e-step update given by Equation (15), the denominator sometimes became numerically zero, causing a division-by-zero error. When this occurred, we added 1.0e-12 to the denominator.

Table 4: Datasets used in experiments for PMF estimation

	# Feature D	# Non-zero values N	Tensor size $\lvert\Omega\rvert$	Sparsity $N/\lvert\Omega\rvert$	# Classes I_D
AsiaLung	8	40	2.56e+02	1.56e-01	2
BalanceScale	5	437	1.88e+03	2.33e-01	3
CarEvaluation	7	1209	6.91e+03	1.75e-01	4
Chess2	7	19639	1.18e+06	1.66e-02	18
Cleveland	8	139	5.76e+03	2.41e-02	5
ConfAd	7	129	5.88e+03	2.19e-02	2
Coronary	6	61	6.40e+01	9.53e-01	2
Credit	10	309	1.09e+05	2.84e-03	2
DMFT	5	425	2.27e+03	1.87e-01	6
GermanGSS	6	280	8.00e+02	3.50e-01	2
Hayesroth	5	61	5.76e+02	1.06e-01	3
Led7	8	281	1.28e+03	2.20e-01	10
Lenses	4	8	2.40e+01	3.33e-01	3
Mofn	11	777	2.05e+03	3.79e-01	2
Monk	7	302	8.64e+02	3.50e-01	2
Nursery	9	9072	6.48e+04	1.40e-01	5
PPD	9	53	7.78e+03	6.82e-03	2
PTumor	16	167	9.83e+04	1.70e-03	2
Parity5p5	11	735	2.05e+03	3.59e-01	2
Sensory	12	403	4.42e+05	9.11e-04	2
ThreeOfNine	10	358	1.02e+03	3.50e-01	2
Tumor	13	195	1.29e+05	1.51e-03	21
Vehicle	5	33	9.60e+01	3.44e-01	2
Votes	17	123	1.31e+05	9.38e-04	2
XD6	10	455	1.02e+03	4.44e-01	2

We emphasize that even if the χ-divergence $D_\chi(\mathbf{T}, \mathbf{P})$ itself is a convex function with respect to the tensor $\mathbf{P}$, directly applying the Newton method to the many-body approximation without using the θ- and η-coordinate system does not guarantee a globally optimal solution since the factorization constraint on the model manifold $\mathscr{B}_\chi$ is non-convex.

B EXPERIMENTAL DETAILS

Setup for PMF estimation We follow standard evaluation methodologies for tensor-based PMF estimation used in (Glasser et al., 2019; Kargas et al., 2018). Given discrete samples with D features, $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(N)}$, where each $\boldsymbol{x}^{(n)} \in \Omega = [I_1] \times \cdots \times [I_D]$, we divide the samples into a training dataset $\mathbf{D}^{\text{train}}$, a validation dataset $\mathbf{D}^{\text{valid}}$, and a test dataset $\mathbf{D}^{\text{test}}$. Correspondingly, we construct a training tensor $\mathbf{T}^{\text{train}} \in \mathbb{R}^{I_1 \times \cdots \times I_D}$, a validation tensor $\mathbf{T}^{\text{valid}} \in \mathbb{R}^{I_1 \times \cdots \times I_D}$, and a test tensor $\mathbf{T}^{\text{test}} \in \mathbb{R}^{I_1 \times \cdots \times I_D}$. Specifically, we set $\mathbf{T}^{\ell}_{\boldsymbol{i}} = 1/N^{\ell} \sum_{n=1}^{N} \delta(\boldsymbol{x}^n, \boldsymbol{i})$, where N^ℓ denotes the number of samples in $\mathbf{D}^\ell$ for $\ell \in \{\text{train}, \text{valid}, \text{test}\}$, and $\delta(\boldsymbol{x}^n, \boldsymbol{i})$ is the Kronecker delta, i.e., $\delta(\boldsymbol{x}^n, \boldsymbol{i}) = 1$ if $\boldsymbol{x}^n = \boldsymbol{i}$, 0 otherwise. Thus, the tensor $\mathbf{T}^{\text{train}}$ can be naturally regarded as an empirical distribution. We then decompose $\mathbf{T}^{\text{train}}$ using either the proposed or baseline methods to obtain the reconstruction $\mathbf{P}$. Model hyperparameters such as the (deformed) ranks, bodies, κ, and q are tuned to maximize the validation score, defined as the log-likelihood. Finally, we evaluate the test log-likelihood $\sum_{\boldsymbol{i} \in \Omega} \mathbf{T}^{\text{test}}_{\boldsymbol{i}} \log \mathbf{P}_{\boldsymbol{i}}$, where the tensor $\mathbf{P}$ is the reconstruction of $\mathbf{T}^{\text{train}}$ with the tuned hyperparameters. It should be noted that the proposed methods consistently ensure normalization in the standard summation sense, i.e., $\sum_{\boldsymbol{i}} \mathbf{P}_{\boldsymbol{i}} = 1$, thereby guaranteeing that

Table 5: Datasets source and license information.

Dataset	License	URL
AsiaLung	Public	https://www.openml.org/search?type=data&status=active&id=43151
BalanceScale	CC BY 4.0	https://archive.ics.uci.edu/dataset/12/balance+scale
CarEvaluation	CC BY 4.0	https://archive.ics.uci.edu/dataset/19/car+evaluation
Chess2	CC BY 4.0	https://archive.ics.uci.edu/dataset/23/chess+king+rook+vs+king
Cleveland	Public	https://www.openml.org/search?type=data&status=active&id=40711
ConfAd	CC0	https://www.openml.org/search?type=data&status=active&id=41538
Coronary	Public	https://www.openml.org/search?type=data&status=active&id=43154
Credit	CC BY 4.0	https://archive.ics.uci.edu/dataset/27/credit+approval
DMFT	Public	https://www.openml.org/search?type=data&status=active&id=469
GermanGSS	Public	https://www.openml.org/search?type=data&status=any&id=1025&sort=runs
Hayesroth	CC BY 4.0	https://archive.ics.uci.edu/dataset/44/hayes+roth
Led7	Public	https://www.openml.org/search?type=data&status=active&id=40678&sort=runs
Lenses	CC BY 4.0	https://archive.ics.uci.edu/dataset/58/lenses
Mofn	Public	https://www.openml.org/search?type=data&sort=runs&id=40680&status=active
Monk	CC BY 4.0	https://archive.ics.uci.edu/dataset/70/monk+s+problems
Nursery	CC BY 4.0	https://archive.ics.uci.edu/dataset/76/nursery
PPD	Public	https://www.openml.org/search?type=data&status=any&id=40683
PTumor	Public	https://www.openml.org/search?type=data&status=active&id=1003
Parity5p5	MIT	https://epistasislab.github.io/pmlb/profile/parity5+5.html
Sensory	Public	https://www.openml.org/search?type=data&sort=runs&id=826&status=active
ThreeOfNine	Public	https://www.openml.org/search?type=data&status=active&id=40690
Tumor	CC BY 4.0	https://archive.ics.uci.edu/dataset/83/primary+tumor
Vehicle	Public	https://www.openml.org/search?type=data&status=active&id=835
Votes	CC BY 4.0	https://archive.ics.uci.edu/dataset/105/congressional+voting+records
XD6	Public	https://www.openml.org/search?type=data&status=active&id=40693

the model produces a valid density and the evaluation based on the log-likelihood is fair. [1] Each sample in the datasets used in these experiments has a class label. We therefore also evaluate the obtained density $\mathbf{P}$ by its classification performance, which can be regarded as a conditional PMF estimation task. Specifically, each sample is classified according to the class label that maximizes the conditional probability of the class label. In this experiment, we tune the hyperparameters to maximize the classification accuracy on the validation dataset.

Dataset selection criteria for PMF estimation For the PMF estimation experiments where we evaluate the performance of the proposed deformed many-body approximation, we used 25 discrete datasets from the UCI machine learning repository[2] and the PMLB benchmark[3](Olson et al., 2017; Romano et al., 2021), each with a tensor size (i.e., sample space size $|\Omega|$) of less than one million, since the methods require access to all tensor elements for parameter updates in Equation (14) and are therefore not scalable to huge datasets. For the additional experiments in Appendices C.1 and C.3 involving more computationally demanding runs, we selected seven datasets from the above 25 that have tensor sizes below 50,000 and at least 250 samples resulting in the following datasets: BalanceScale, CarEvaluation, GermanGSS, Mofn, Monk, ThreeOfNine, and XD6. The used datasets are summarized in Tables 4 and 5 with license information and dataset sources.

Dataset detail for the COIL image For the image reconstruction experiments, we selected five color images (object numbers 4, 5, 7, 10, and 17) from the COIL dataset (Nayar and Murase, 1996) used in the baseline work (Ghalamkari et al., 2023). In the original work in (Ghalamkari et al., 2023), additional images (object numbers 23, 28, 38, 42, and 78) were also stacked and resized to 40×40, resulting in a $40 \times 40 \times 3 \times 10$ tensor for

[1] In deformed algebra, there is an operation called the deformed sum $\oplus$ which is defined as $\mathrm{Log}_\chi[a] \oplus \mathrm{Log}_\chi[b] = \mathrm{Log}_\chi[ab]$ (Wada and Scarfone, 2023). However, we did not ensure the normalization of tensors in the sense of this deformed summation, as it does not guarantee a valid density.

[2] https://archive.ics.uci.edu/

[3] https://epistasislab.github.io/pmlb/

decomposition. In contrast, in this study, we reduced the number of stacked images and increased the image resolution to 60×60 to more precisely investigate the reconstruction performance under noise. The input images are shown in Figure 12. The dataset is publicly available from the official website.[4]. The tensor $\mathbf{T}$ was normalized by dividing it by its total sum, and the proposed method was then applied. In the experiments with additive noise, we randomly selected n entries and replaced their values with either 0 or 1 with equal probability, which is often referred to as salt-and-pepper noise (Chanu et al., 2023).

Dataset detail for the synthetic data For the experiments with synthetic data, we constructed tensors of rank at most 100 from non-negative factor matrices $A^{(1)}, A^{(2)}, \dots, A^{(D)} \in \mathbb{R}^{n \times 100}$, where each entry was sampled independently from the uniform distribution on $[0, 1]$. The resulting tensor $\mathbf{T} \in \mathbb{R}^{n \times \cdots \times n}$ was then normalized by dividing all entries by their total sum.

Detailed setup for CPGDs We tuned rank within eight equally spaced integers so that the number of parameters in the CP-model does not exceed half of the number of samples, which is the same policy as experiments with baselines. The iterations were stopped at 10000 iterations. The learning rate was tuned within [1.0e4, 1.0e3, ..., 1.0e-4] to maximize the validation score. Since CPGD1 and CPGD2 have initial value dependency, we run each algorithm five times with random initialization and evaluate the mean and standard deviation.

Summary of baselines Since this study focuses on investigating the advantages provided by the deformed product, only tensor decomposition-based methods were included as baselines in all experiments. Specifically, we included the traditional EMCP, which optimizes the KL divergence via the EM algorithm and has been shown to perform well in PMF estimation (Huang and Sidiropoulos, 2017; Yeredor and Haardt, 2019; Chege et al., 2022). We also included KL-divergence-based methods for PMF estimation with more complex decomposition structures, such as Born Machine (BM) and Tensor Train decomposition (KLTT) (Glasser et al., 2019). Additionally, we included E^2MCPTTB, which enhances generalization through tensor mixtures and α-divergence minimization (Ghalamkari et al., 2024), and CNMF (Ibrahim and Fu, 2021), which adapts coupled tensor decomposition techniques. To isolate the benefits of the deformed algebra from those arising from α-divergence optimization, we also included q-CP based on the E^2MCP algorithm (Ghalamkari et al., 2024), which performs CP decomposition via α-divergence optimization. Note that Tucker decomposition based on multiplicative updates (Kim et al., 2008) was not considered as a baseline, as normalization is not guaranteed.

Additional information for reproducibility All experiments were conducted on a university HPC cluster without GPU acceleration. The compute nodes run Linux kernel 6.1.149 on `x86_64` architecture and are equipped with 251 GB of RAM. Python 3.12.7 was used for all computations, and jobs were submitted using `bsub`. During computation, memory usage did not exceed 19 GB. For reproducibility, the source code used in our experiments is included in the supplementary material.

C ADDITIONAL NUMERICAL RESULTS

Due to space limitations, we discuss the results of additional experiments here.

C.1 PMF estimation on real datasets

The full version of Table 1 is presented in Table 7, covering all 25 datasets. Overall, despite the advantage of being free from initialization dependence and global optimization, the proposed methods, q- and κ-MBA, achieve performance comparable to conventional non-convex methods. In Table 7, we also report the optimal values of body, q, and κ that maximize the validation scores, namely the classification accuracy and log-likelihood. Interestingly, the parameters that yield the best classification performance are often different from the $\kappa = 0$, suggesting that minimizing the KL divergence is not necessarily the most suitable objective for classification. In contrast, the optimal q for maximizing the log-likelihood frequently turns out to be $q = 1$, which is expected since negative log-likelihood minimization is equivalent to optimizing the KL divergence. On the other hand, and somewhat unexpectedly, the optimal κ that maximizes the log-likelihood is often far from 0. This indicates that, for maximizing the test log-likelihood, optimizing the KL divergence may not always be the appropriate choice.

[4] `https://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php`

To further investigate the properties of q-MBA and κ-MBA, we randomly changed the class labels of m-samples in the training and validation data, and examined how the quality of the estimated mass varies with q and κ. The results are shown in Figure 10. As the number of mislabeled samples m increases, the quality of the estimated mass becomes worse. However, for κ-MBA, the increase in test negative log-likelihood becomes less steep. This behavior was not observed for q-MBA.

C.2 Additive noise robustness in Tsallis-deformed many-body approximation

Although q-MBA optimizes the Tsallis divergence, its optimization is equivalent to that of the α-divergence, which provides robustness to noise, as discussed in Section 3.1. Therefore, q-MBA is expected to inherit this noise robustness. To verify this, we extracted five images from the COIL dataset to construct a $60\times60\times3\times5$ tensor, added salt-and-pepper noise, and performed the Tsallis-deformed three-body approximation. The resulting reconstructions are shown in Figure 12. We also provide Figure 6(left), which shows how the quality of reconstruction depends on the parameter q and the noise level. It can be observed that larger values of q result in less sensitive reconstructions to noise. This observation is consistent with the known fact that α-divergence induces mass-covering behavior when α is smaller (Li and Farnia, 2023; Hernandez-Lobato et al., 2016). The results of the same experiments by κ-MBA are shown in Figures 13 and 6(right); however, the stability against additive noise observed in q-MBA was not found in κ-MBA. We provide an explanation of these results from the sensitivity of the model to perturbations in Appendix G.1.

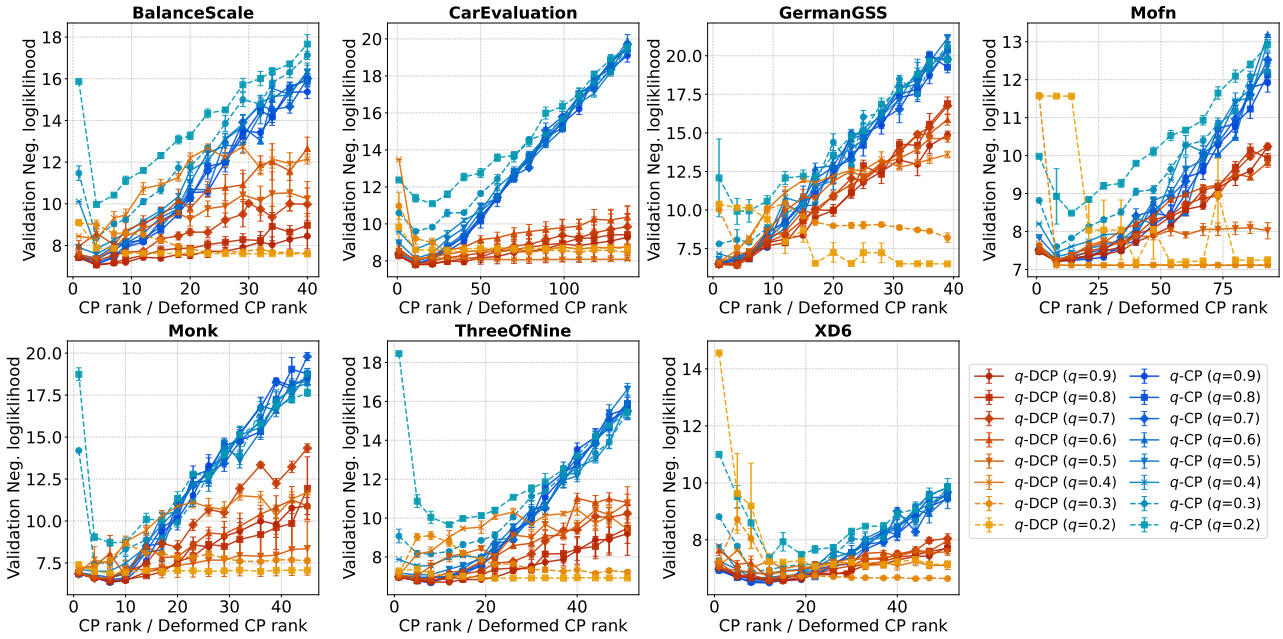
We note that q-MBA and κ-MBA are identical to the ordinary KL-divergence-based many-body approximation in the limits $q \to 1$ and $\kappa \to 0$, respectively. In fact, the experimental results show that they converge to the same curve in these limits.

C.3 Regularization induced by Tsallis-deformed algebra

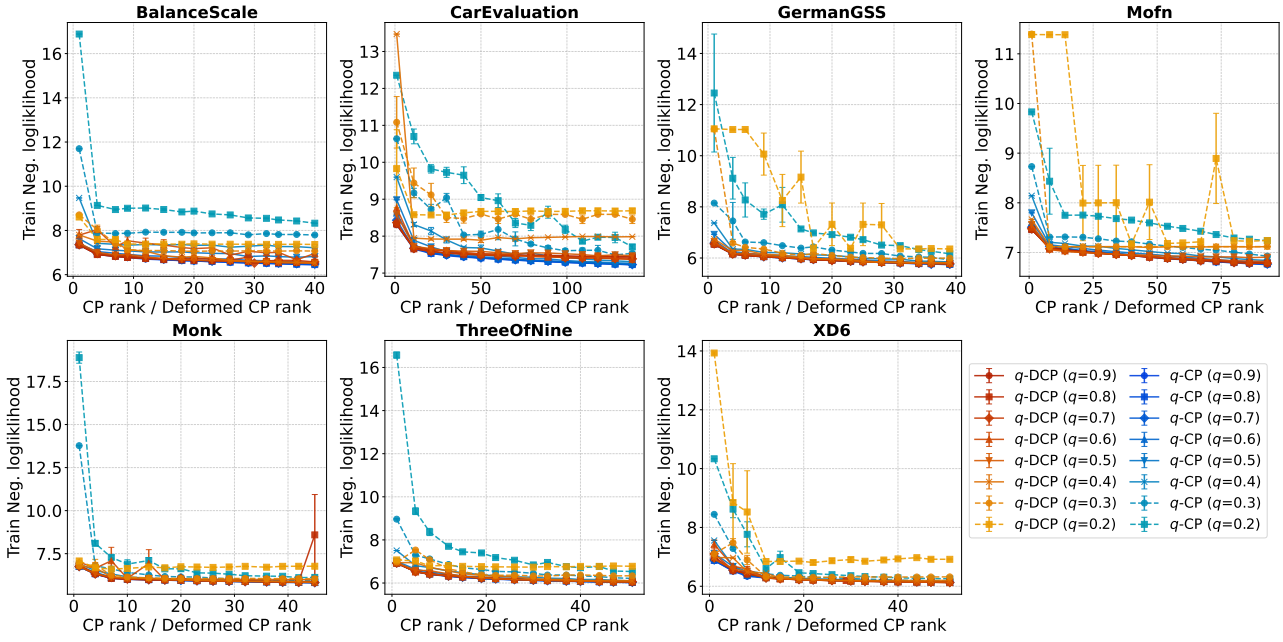
The full results of Figure 2(**a**) for varying values of q and across different datasets are shown in Figure 11. In all experiments, the (deformed) rank of the low-rank model was chosen such that the number of parameters in the low-rank model approximately matches that of the three-body approximation. The results show that the proposed q-MBA achieves lower training error than q-CP in most datasets. Theoretically, optimization based on L-BFGS should find the same solution as Newton-based optimization. However, when q is small, L-BFGS often becomes unstable, occasionally triggering convergence prematurely, as seen in the results for the CarEvaluation datasets. Although we attempted to address this numerical instability by adjusting the convergence threshold, we could not fully resolve it. The experimental results on the XD6 dataset with $q = 0.3$ show that L-BFGS achieved slightly better optimization. According to the optimization log, the Newton method issued a line search failure warning just before satisfying the convergence criterion. When q is small, the q-DCP loses model capacity due to the regularization effect, preventing the training error from being sufficiently reduced, as discussed in Section 3.1.

To examine the regularization performance of the proposed q-DCP, we also provide the full results of Figure 2(**b**) across different datasets. Note that both q-CP and q-DCP optimize the α-divergence using an EM-like algorithm; the main difference is that q-DCP relies on Tsallis-deformed algebra. Therefore, this comparison serves as an appropriate ablation study to isolate the effect of the deformed algebra. The results in Figures 4 and 5 show that q-CP is more prone to overfitting the training data when the rank is larger, whereas q-DCP is not, which is consistent with the regularization property of q-DCP supported by Proposition 7. We note that solutions of the q-DCP and q-CP depend on the initial values and may converge to a stationary point, while the q-MBA and κ-MBA can obtain the global optimum regardless of initialization. Thus, each experiment for q-CP and q-DCP was repeated 5 times with random initializations.

The regularization effect induced by Tsallis-deformed algebra can also be observed in the noisy image reconstruction task. We obtained a non-negative color image $\mathbf{T}^{\text{true}} \in \mathbb{R}^{60\times60\times3}$ from the COIL dataset and added salt-and-pepper noise to construct a noisy training tensor $\mathbf{T}^{\text{noisy}} \in \mathbb{R}^{60\times60\times3}$. The tensor $\mathbf{T}^{\text{noisy}}$ was then decomposed using q-CP and q-DCP to obtain a reconstruction $\mathbf{P} \in \mathbb{R}^{60\times60\times3}$, while varying the (deformed) rank and the value of q. We evaluated the training relative error $\|\mathbf{T}^{\text{noisy}} - \mathbf{P}\|_F / \|\mathbf{T}^{\text{noisy}}\|_F$ and the test relative error $\|\mathbf{T}^{\text{true}} - \mathbf{P}\|_F / \|\mathbf{T}^{\text{true}}\|_F$, as shown in Figure 7. The training relative error decreases as the rank increases, whereas the model also fits the noise in the training data, leading to larger test relative errors, which is a typical bias-variance tradeoff. However, when the value of q is small, the decrease in training error is suppressed since the model has no capacity as supported by Proposition 7. Additionally, the test error curve plateaus for certain values of q, offering stable

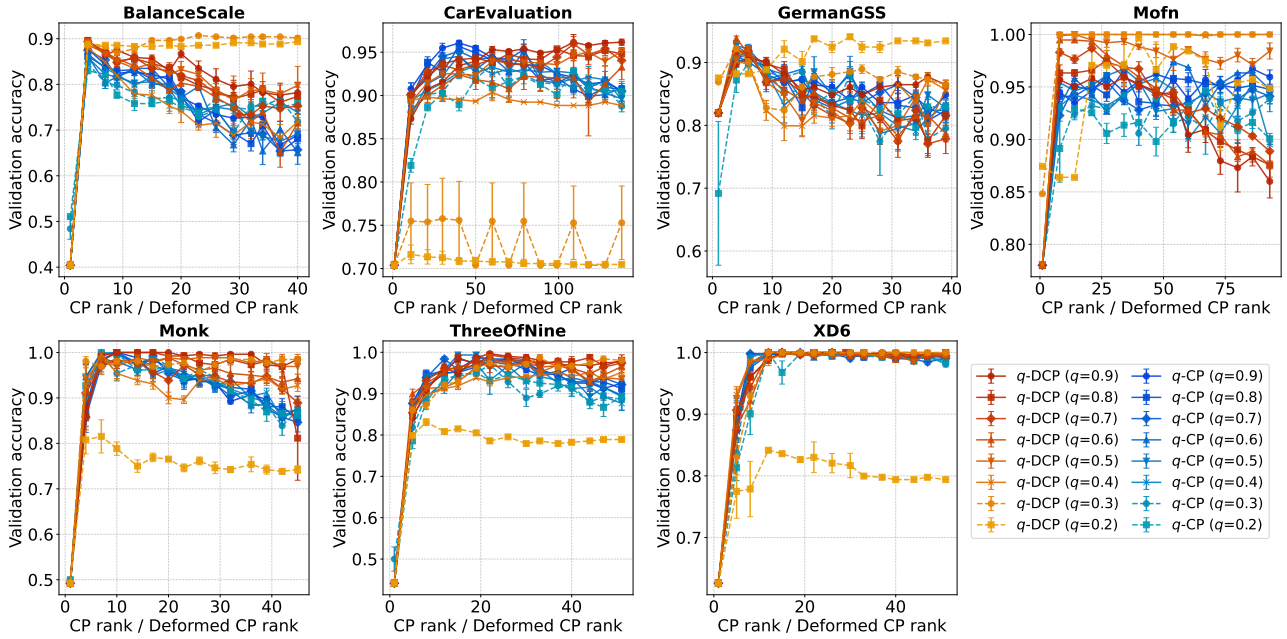


(a) Negative log-likelihood for validation data

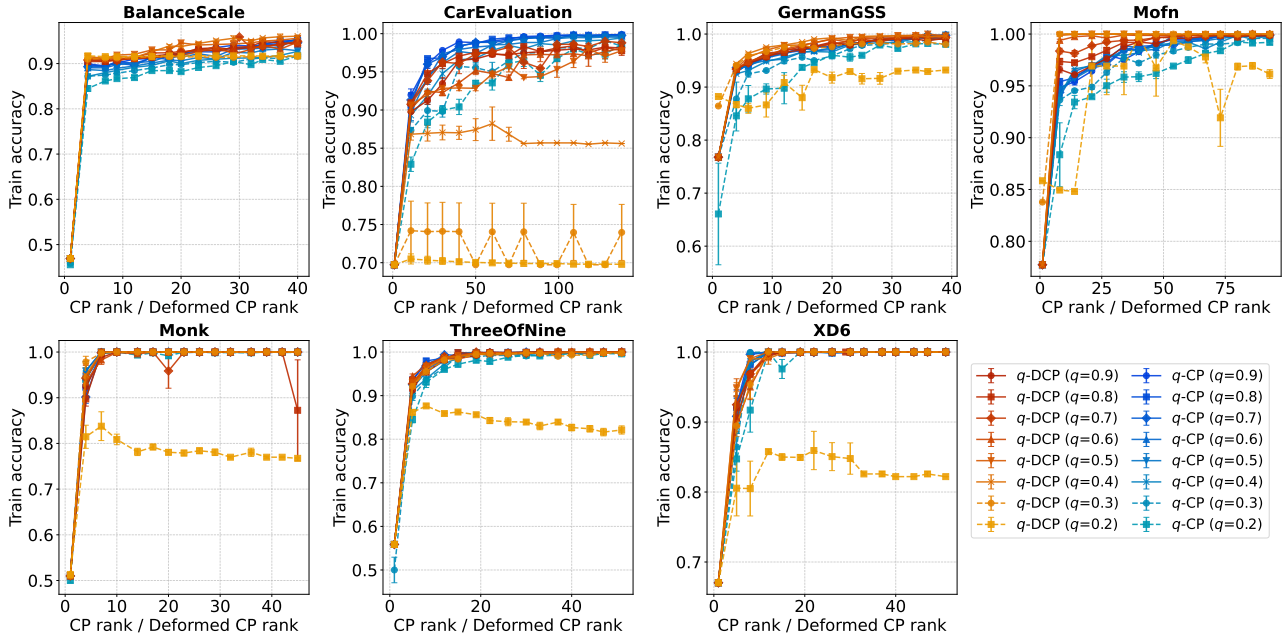


(b) Negative log-likelihood for training data

Figure 4: Negative log-likelihood (lower is better) of q-CP and q-DCP for different (deformed) ranks and q values on validation (a) and training (b) data. Each experiment was repeated 5 times with random initializations, and the mean values are plotted with error bars representing the standard error of the mean (SEM).



(a) Classification accuracy for validation datasets



(b) Classification accuracy for training datasets

Figure 5: Classification accuracy (higher is better) of q-CP and q-DCP for different (deformed) ranks and q values on validation (a) and training (b) data. Each experiment was repeated 5 times with random initializations, and the mean values are plotted with error bars representing the standard error of the mean (SEM).

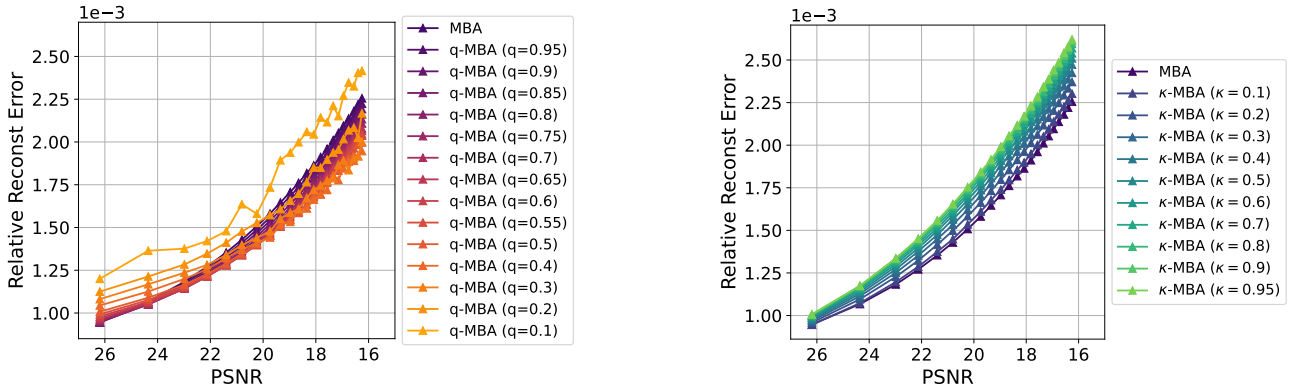


Figure 6: Relative reconstruction error of noisy COIL images by the q- and κ-deformed three-body approximation. The reconstructed images can be found in Figure 12.

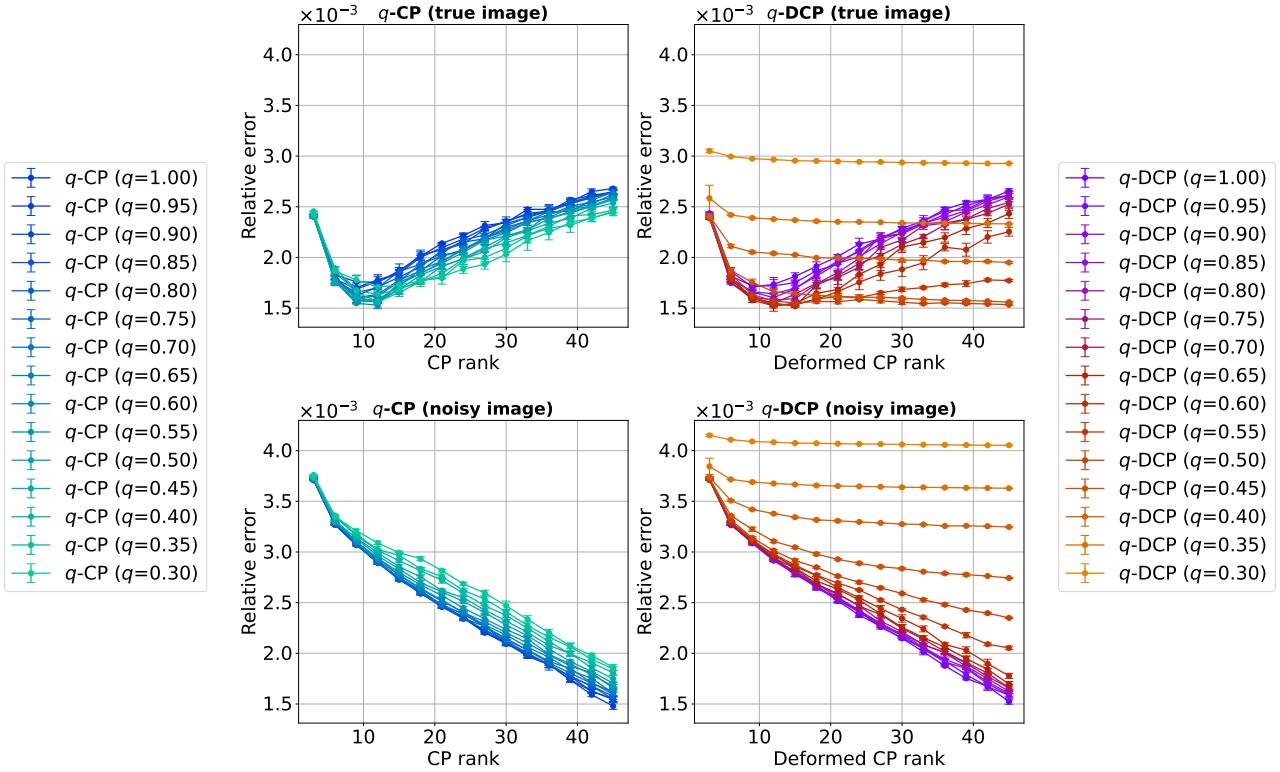


Figure 7: Test error (top panels) and training error (bottom panels) of noisy image learning by q-CP (left panels) and q-DCP (right panels). Each experiment was repeated 5 times with random initializations, and the mean values are plotted with error bars representing the standard error of the mean (SEM).

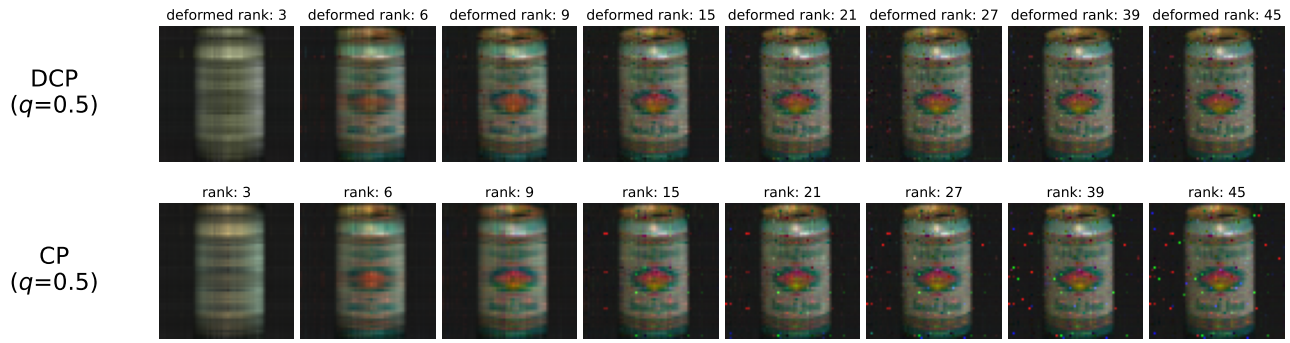


Figure 8: Reconstructions of a noisy input image using q-CP and q-DCP ($q = 0.5$) with varying (deformed) rank. The reconstruction by q-DCP effectively suppresses noise compared to q-CP.

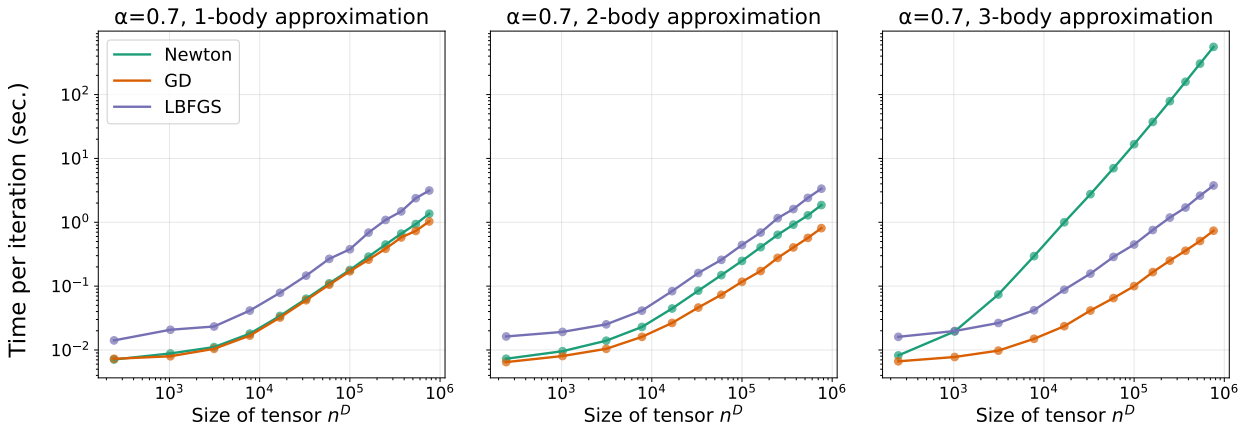


Figure 9: Running time per iteration of the Tsallis-deformed m-body approximation on a random synthetic tensor of size (n, n, n, n, n) using three optimization methods: Newton, first-order gradient descent (GD), and L-BFGS. We measured the time required for 10 iterations and plotted the average value per iteration.

learning. These results suggest that an appropriately chosen q achieves better generalization by preventing overfitting, due to the regularization induced by the deformed algebra. In Figure 8, we visualize the reconstructed tensor $\mathbf{P}$ for $\alpha = 0.5$ using q-CP and q-DCP. The reconstruction produced by q-DCP shows effective noise suppression, while q-CP tends to fit the noise when the rank is increased.

We also note that the q-DCP and q-CP become algorithmically identical in the limit as $q \to 1$. Consistently, the experimental results in this limit asymptotically converge to the same curve.

C.4 Computational time

Figure 9 shows the computation time per iteration for each of these optimization methods in the Tsallis-deformed m-body approximation, while varying the size of the input tensor. Although the methods that utilize the χ-Fisher information matrix $\mathbf{G}$ require a higher computational cost compared to GD, it can be observed that L-BFGS reduces this cost.

C.5 Comparison to Frobenius norm minimization in PMF estimation

To verify that the Frobenius norm optimization is, in general, inferior to the KL divergence optimization for PMF estimation, we provide additional experimental results. As a fair comparison, we provide below the direct comparison of the low-rank CP model by EM-based KL-divergence optimization (EMCP) to Frobenius norm optimization by the gradient-based method (CPGD). We implemented two versions of CPGDs:

- CPGD1, which applies a softmax function to the factor matrices and weights of each rank-1 component to

Table 6: Experimental results for PMF estimation comparing Frobenius norm minimization, evaluated in negative log-likelihood (left block) and classification accuracy (right block).

	Negative log-likelihood			Classification accuracy		
dataset	CPGD1	CPGD2	EMCP	CPGD1	CPGD2	EMCP
AsiaLung	2.27(0.01)	2.48(0.13)	**2.26(0.00)**	**0.57(0.00)**	**0.57(0.00)**	**0.57(0.00)**
BalanceScale	**7.16(0.02)**	**7.16(0.01)**	8.98(0.49)	0.85(0.01)	**0.86(0.01)**	**0.86(0.01)**
CarEvaluation	**7.77(0.04)**	8.34(0.00)	7.78(0.04)	**0.95(0.01)**	0.71(0.00)	0.94(0.02)
Chess2	12.49(0.04)	13.97(0.03)	**11.95(0.08)**	0.36(0.02)	0.09(0.06)	**0.44(0.01)**
Cleveland	6.24(0.00)	6.34(0.02)	**6.14(0.00)**	0.49(0.00)	0.55(0.05)	**0.58(0.00)**
ConfAd	7.13(0.05)	7.03(0.02)	**6.85(0.00)**	**0.91(0.00)**	**0.91(0.00)**	**0.91(0.00)**
Coronary	**3.51(0.00)**	3.52(0.01)	**3.51(0.00)**	**0.64(0.00)**	**0.64(0.00)**	**0.64(0.00)**
Credit	**7.15(0.01)**	10.43(0.23)	8.37(0.00)	**0.92(0.00)**	0.54(0.08)	0.55(0.00)
DMFT	7.23(0.03)	7.33(0.04)	**7.17(0.00)**	**0.23(0.02)**	0.18(0.02)	**0.23(0.00)**
GermanGSS	6.35(0.00)	6.46(0.09)	**6.31(0.00)**	**0.83(0.02)**	**0.83(0.03)**	**0.83(0.01)**
Hayesroth	6.33(0.11)	6.22(0.04)	**5.80(0.05)**	0.26(0.04)	0.30(0.10)	**0.69(0.03)**
Led7	4.69(0.04)	6.21(0.03)	**4.65(0.03)**	0.43(0.00)	0.17(0.04)	**0.45(0.00)**
Lenses	3.01(0.04)	3.15(0.06)	**2.94(0.00)**	**0.75(0.00)**	**0.75(0.00)**	**0.75(0.00)**
Mofn	7.27(0.01)	7.34(0.05)	**7.21(0.01)**	0.96(0.01)	0.83(0.03)	**0.97(0.00)**
Monk	**6.28(0.06)**	6.65(0.00)	6.31(0.09)	**1.00(0.00)**	0.72(0.00)	**1.00(0.00)**
Nursery	9.75(0.01)	10.81(0.03)	**9.64(0.02)**	0.92(0.01)	0.33(0.01)	**0.99(0.00)**
PPD	7.63(0.00)	7.65(0.00)	**7.04(0.00)**	**0.61(0.00)**	**0.61(0.00)**	**0.61(0.00)**
PTumor	7.79(0.23)	9.35(0.98)	**7.39(0.00)**	0.66(0.01)	0.66(0.00)	**0.74(0.01)**
Parity5p5	**7.21(0.11)**	7.63(0.01)	7.43(0.03)	**0.99(0.01)**	0.50(0.03)	0.95(0.04)
Sensory	12.66(0.14)	13.03(0.01)	**11.72(0.08)**	0.68(0.00)	0.59(0.16)	**0.72(0.00)**
ThreeOfNine	6.74(0.02)	6.77(0.03)	**6.66(0.00)**	**0.96(0.02)**	0.81(0.02)	**0.96(0.01)**
Tumor	7.91(0.01)	10.37(0.63)	**7.74(0.00)**	0.29(0.04)	0.21(0.03)	**0.30(0.00)**
Vehicle	4.64(0.08)	4.62(0.05)	**4.52(0.00)**	0.71(0.00)	0.89(0.14)	**1.00(0.00)**
Votes	8.32(0.05)	11.70(0.04)	**7.87(0.00)**	0.81(0.01)	0.43(0.19)	**0.84(0.00)**
XD6	6.50(0.01)	6.73(0.01)	**6.49(0.01)**	**1.00(0.00)**	0.80(0.02)	**1.00(0.00)**

ensure both non-negativity and normalization,

- CPGD2, which applies a softplus function to the factor matrices to guarantee non-negativity only, while optimizing the Frobenius norm of the difference between the input tensor $\mathbf{T}$ and the normalized low-rank tensor $\mathbf{P}$, i.e., minimizing $\|\mathbf{T} - \mathbf{P}/\sum_{\boldsymbol{i}} \mathbf{P}_{\boldsymbol{i}}\|_F^2$.

We note that the proposed MBA and baseline methods in Table 1 automatically satisfy the non-negative normalized conditions. We conducted PMF estimation in the same setting as described in Appendix B. The results are presented in Table 6, demonstrating that the KL divergence optimization, in general, has better performance than the Frobenius norm optimization.

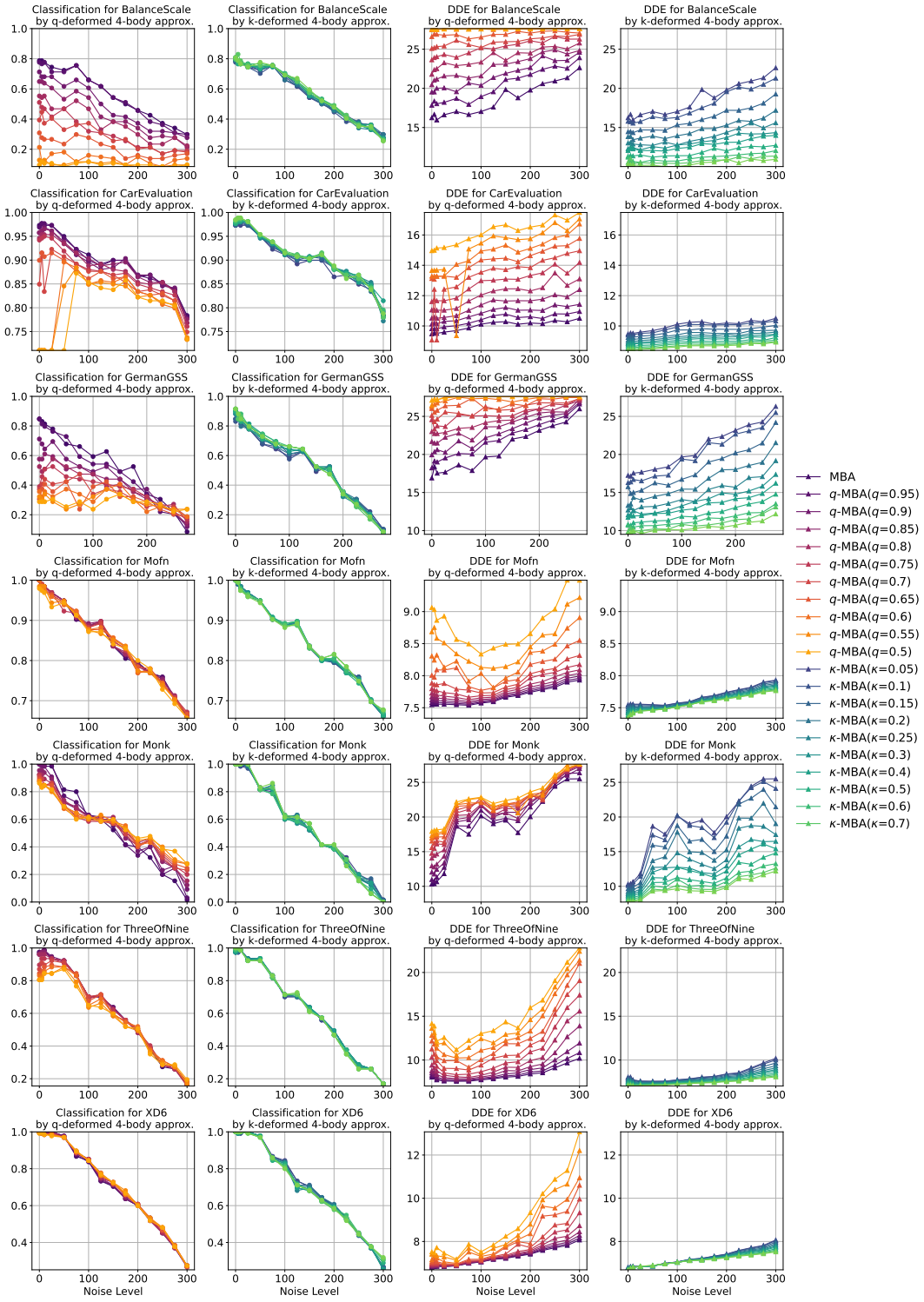


Figure 10: Results of PMF estimation using the Tsallis-deformed and Kaniadakis-deformed 4-body approximations. The vertical axis shows accuracy (higher is better) and negative log-likelihood (lower is better) on test samples, while the horizontal axis indicates the number of samples with randomly permuted class labels in the training dataset. The experimental results asymptotically converge to the curve of the existing KL-divergence-based many-body approximation (MBA) in the limits as $q \to 1$ and $\kappa \to 0$.

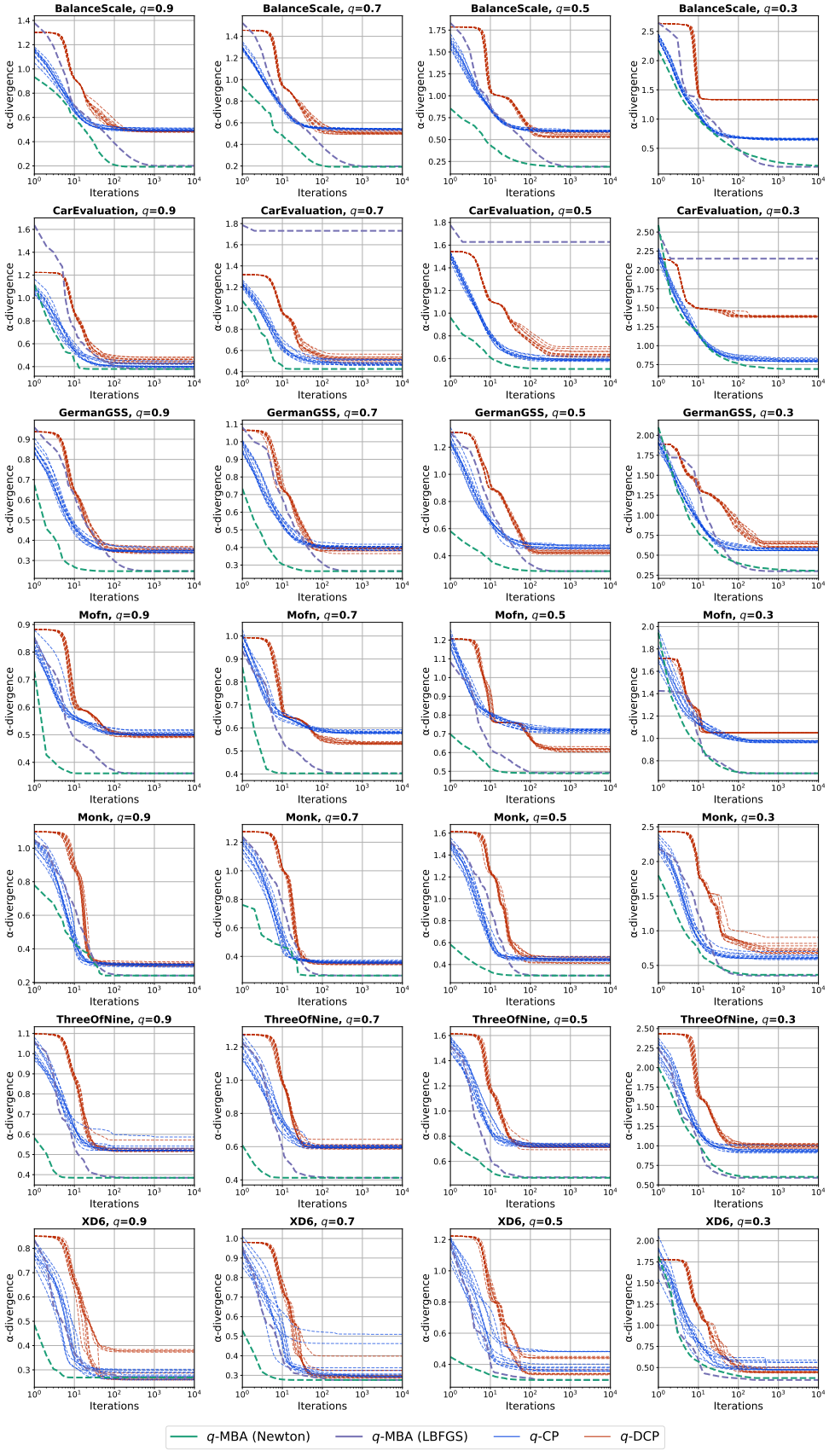


Figure 11: Training error per iteration for q-CP, q-DCP, and q-MBA using Newton and L-BFGS methods. Since both q-DCP and q-CP exhibit dependence on initialization, we performed 10 runs with random initializations and plotted all the results. In contrast, q-MBA is free from initialization dependence, and consequently, only a single run was plotted.

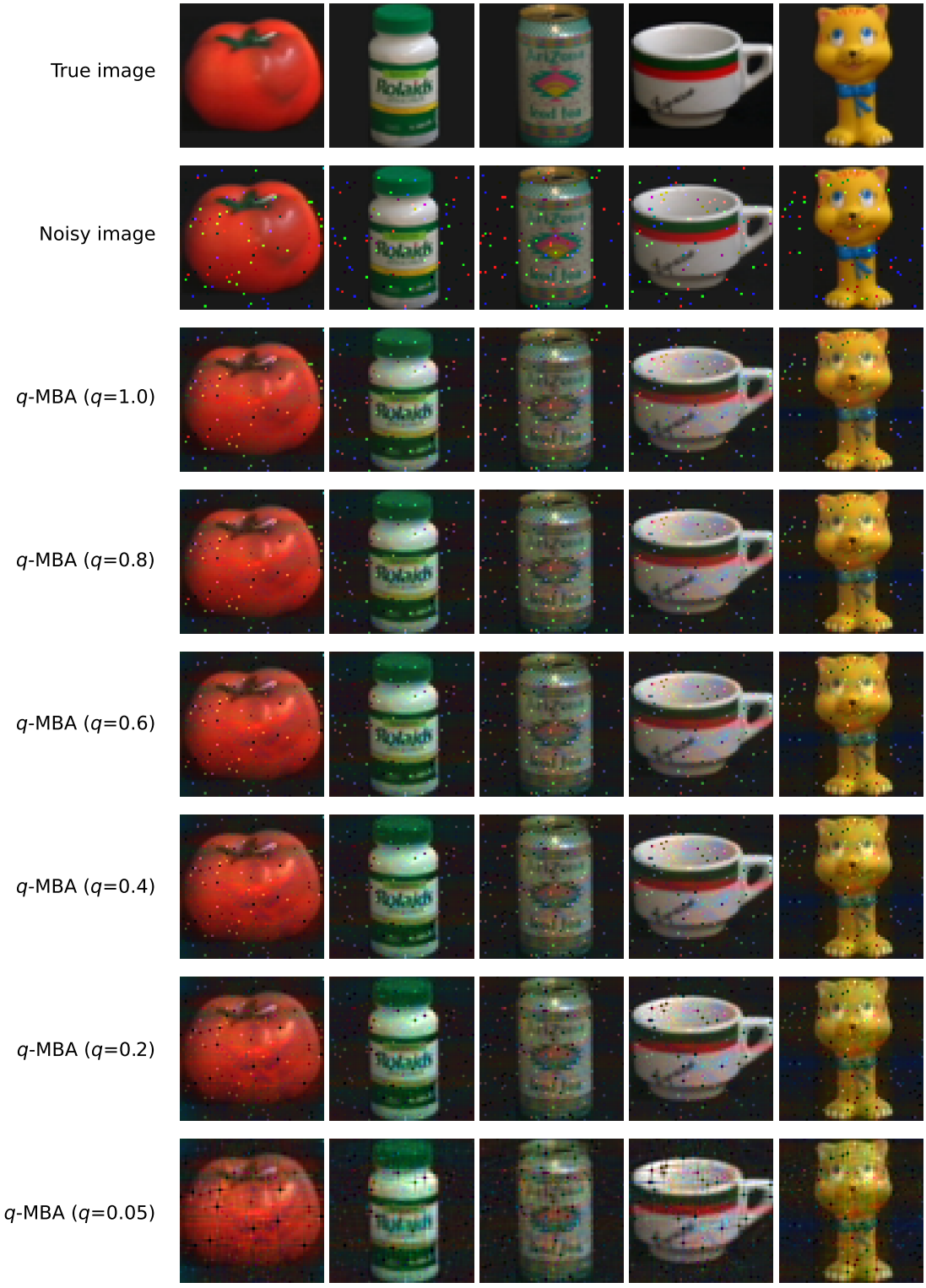


Figure 12: Reconstruction of noisy COIL images (PSNR = 22) by Tsallis-deformed three-body approximations (q-MBA). The Tsallis-deformed approximation with $q = 1.0$ corresponds to the ordinary KL-divergence-based many-body approximation.

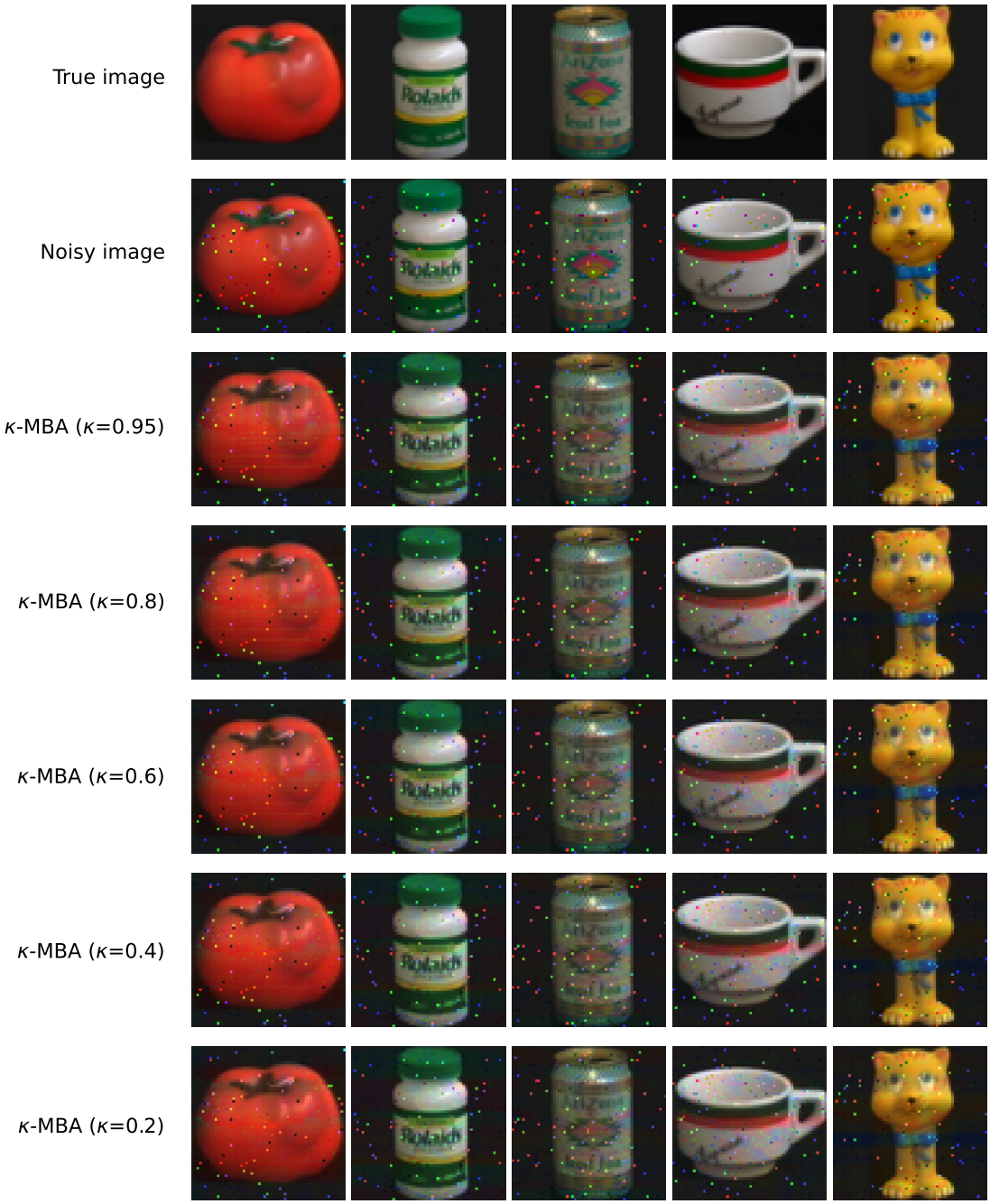


Figure 13: Reconstruction of noisy COIL images (PSNR = 22) by Kaniadakis-deformed three-body approximations (κ-MBA).

Table 7: Accuracy of the classification (top lines, higher is better) and negative log-likelihood (bottom lines, lower is better) for the 25 datasets. Error bars for baselines are given as the standard deviation of the mean across 5 randomly initialized runs for baselines. Since the proposed q-MBA and κ-MBA have no initial value dependency, we report results from a single run. For these proposed methods, the optimal body, q, and κ values maximizing the validation score are also reported.

Dataset	BM	KLTT	CNMF	EMCP	E^{2}MCPTTB	q-MBA [q,body]	κ-MBA [κ,body]
AsiaLung	0.57 (0.00) **2.24** (0.00)	**0.59** (0.02) 2.29 (0.06)	0.57 (0.00) 3.13 (0.09)	0.57 (0.00) 2.26 (0.00)	0.50 (0.00) 2.48 (0.16)	0.57 [1.0, 2] 2.27 [1.0, 2]	0.57 [0.0, 2] 2.25 [0.7, 3]
BalanceScale	0.78 (0.00) 7.32 (0.03)	0.83 (0.00) 7.48 (0.15)	0.65 (0.07) 7.42 (0.02)	0.86 (0.01) 8.98 (0.49)	0.87 (0.00) 7.10 (0.00)	**0.89** [1.0, 2] 7.04 [1.0, 2]	**0.89** [0.0, 2] **7.03** [0.3, 2]
CarEvaluation	0.85 (0.04) 7.96 (0.14)	0.89 (0.00) **7.75** (0.02)	0.74 (0.01) 8.19 (0.06)	0.94 (0.02) 7.78 (0.04)	0.95 (0.02) 7.85 (0.07)	0.97 [1.0, 4] 7.79 [1.0, 2]	**0.98** [0.15, 4] 7.80 [-0.95, 3]
Chess2	0.16 (0.00) 13.13 (0.00)	0.28 (0.01) 12.56 (0.12)	0.24 (0.01) 12.77 (0.03)	0.44 (0.01) 11.95 (0.08)	0.57 (0.00) 11.64 (0.00)	**0.86** [0.95, 5] 11.42 [1.0, 3]	**0.86** [-0.2, 5] **11.38** [-0.15, 3]
Cleveland	0.54 (0.02) 6.38 (0.05)	**0.58** (0.07) 6.35 (0.27)	0.57 (0.04) **6.14** (0.03)	**0.58** (0.00) **6.14** (0.00)	**0.58** (0.00) 6.24 (0.00)	0.56 [1.0, 2] 6.39 [0.95, 1]	0.49 [0.5, 2] 6.36 [0.8, 2]
ConfAd	0.86 (0.01) 6.92 (0.04)	**0.91** (0.00) 6.98 (0.01)	**0.91** (0.00) 6.92 (0.03)	**0.91** (0.00) 6.85 (0.00)	**0.91** (0.00) 6.86 (0.00)	**0.91** [1.0, 2] 6.85 [1.0, 1]	**0.91** [0.1, 2] **6.83** [0.15, 1]
Coronary	0.63 (0.01) **3.51** (0.00)	**0.64** (0.00) 3.52 (0.00)	**0.64** (0.00) 3.57 (0.01)	**0.64** (0.00) **3.51** (0.00)	**0.64** (0.00) 3.53 (0.00)	0.62 [0.95, 5] 3.52 [1.0, 5]	**0.64** [0.0, 4] 3.53 [0.95, 3]
Credit	0.77 (0.06) 7.12 (0.01)	0.55 (0.00) 7.33 (0.06)	0.68 (0.13) 8.52 (0.17)	0.55 (0.00) 8.37 (0.00)	**0.87** (0.02) 7.03 (0.02)	0.86 [0.65, 2] **6.89** [1.0, 2]	**0.87** [0.9, 2] 6.91 [0.7, 2]
DMFT	0.17 (0.01) 7.31 (0.19)	0.16 (0.03) 7.21 (0.01)	0.20 (0.03) 7.57 (0.09)	**0.23** (0.00) **7.17** (0.00)	0.15 (0.00) 7.21 (0.02)	0.20 [1.0, 2] 7.37 [1.0, 2]	**0.23** [0.7, 2] 7.22 [-0.5, 2]
GermanGSS	0.77 (0.06) 6.56 (0.01)	0.81 (0.04) 6.51 (0.13)	0.79 (0.04) 6.51 (0.04)	0.83 (0.01) **6.31** (0.00)	0.81 (0.00) 6.46 (0.04)	**0.86** [0.55, 2] 6.49 [1.0, 2]	**0.86** [0.8, 2] 6.41 [-0.6, 2]
Hayesroth	0.64 (0.03) 6.64 (0.09)	0.64 (0.03) 6.76 (0.14)	0.31 (0.03) 6.46 (0.19)	0.69 (0.03) **5.80** (0.05)	0.79 (0.17) 5.88 (0.10)	**0.89** [1.0, 2] 6.31 [1.0, 1]	**0.89** [0.0, 2] 6.08 [0.95, 2]
Led7	0.27 (0.04) 5.58 (0.24)	0.19 (0.03) 5.94 (0.24)	0.24 (0.01) 5.74 (0.01)	**0.45** (0.00) 4.65 (0.03)	0.38 (0.00) 4.87 (0.01)	0.44 [1.0, 2] 4.62 [1.0, 2]	0.43 [0.1, 2] **4.61** [0.05, 2]
Lenses	**0.75** (0.00) 2.94 (0.00)	**0.75** (0.00) 2.94 (0.00)	**0.75** (0.00) 3.03 (0.03)	**0.75** (0.00) 2.94 (0.00)	**0.75** (0.00) 3.41 (0.06)	**0.75** [1.0, 1] **2.85** [0.45, 1]	**0.75** [0.0, 1] 2.94 [0.0, 1]
Mofn	**1.00** (0.00) 7.07 (0.01)	0.97 (0.01) **7.05** (0.05)	0.80 (0.00) 7.34 (0.02)	0.97 (0.00) 7.21 (0.01)	**1.00** (0.00) 7.19 (0.01)	**1.00** [1.0, 2] 7.09 [0.55, 2]	**1.00** [0.0, 2] 7.10 [0.0, 2]
Monk	0.98 (0.03) 6.40 (0.05)	0.77 (0.24) 6.65 (0.28)	0.60 (0.07) 6.83 (0.04)	**1.00** (0.00) **6.31** (0.09)	**1.00** (0.01) 6.53 (0.07)	**1.00** [1.0, 3] 6.74 [1.0, 2]	**1.00** [0.0, 3] 6.43 [-0.95, 3]
Nursery	0.73 (0.01) 10.00 (0.04)	0.70 (0.02) 10.02 (0.01)	0.57 (0.07) 10.49 (0.03)	**0.99** (0.00) **9.64** (0.02)	**0.99** (0.00) 9.69 (0.01)	**0.99** [0.4, 3] 9.66 [1.0, 3]	**0.99** [0.0, 4] 9.66 [-0.25, 3]
PPD	0.65 (0.19) 7.07 (0.01)	**0.69** (0.00) 7.07 (0.01)	0.61 (0.00) 7.07 (0.04)	0.61 (0.00) 7.04 (0.00)	0.61 (0.00) 7.52 (0.00)	0.61 [1.0, 1] 7.04 [1.0, 1]	0.61 [0.0, 1] **7.03** [0.05, 1]
PTumor	**0.80** (0.00) 7.77 (0.20)	0.73 (0.00) 7.73 (0.02)	0.66 (0.00) 7.73 (0.04)	0.74 (0.01) **7.39** (0.00)	0.70 (0.00) 7.74 (0.04)	0.75 [0.95, 2] 7.67 [1.0, 1]	0.75 [0.0, 2] 7.75 [0.15, 1]
Parity5p5	0.57 (0.25) 7.54 (0.32)	0.87 (0.23) **7.20** (0.33)	0.48 (0.02) 7.66 (0.01)	0.95 (0.04) 7.43 (0.03)	**1.00** (0.00) 7.26 (0.07)	0.76 [0.85, 6] 7.64 [1.0, 1]	0.72 [0.0, 6] 7.63 [0.95, 1]
Sensory	0.55 (0.01) 11.79 (0.21)	0.66 (0.02) 11.55 (0.00)	0.67 (0.01) 12.61 (0.12)	**0.72** (0.00) 11.72 (0.08)	0.69 (0.02) 11.35 (0.18)	0.63 [1.0, 5] 11.86 [1.0, 2]	0.69 [0.6, 4] **11.17** [-0.9, 3]
ThreeOfNine	0.99 (0.01) 6.64 (0.03)	0.99 (0.01) **6.50** (0.01)	0.76 (0.03) 6.82 (0.01)	0.96 (0.01) 6.66 (0.00)	**1.00** (0.00) 6.72 (0.01)	0.97 [1.0, 4] 6.74 [1.0, 2]	0.97 [0.0, 4] 6.80 [0.95, 3]
Tumor	0.33 (0.00) 7.94 (0.05)	0.24 (0.00) 8.58 (0.42)	0.24 (0.00) 7.81 (0.04)	0.30 (0.00) 7.74 (0.00)	**0.40** (0.03) **7.61** (0.00)	0.37 [0.55, 6] 8.32 [1.0, 1]	0.37 [0.15, 5] 7.99 [0.7, 2]

(continued on next page)

Dataset	BM	KLTT	CNMF	EMCP	E^{2}MCPTTB	q-MBA [q,body]	κ-MBA [κ,body]
Vehicle	0.71 (0.14) 4.99 (0.00)	0.29 (0.14) 4.99 (0.00)	0.93 (0.07) 4.62 (0.10)	**1.00** (0.00) 4.52 (0.00)	**1.00** (0.00) 5.30 (0.15)	**1.00** [1.0, 2] 5.00 [1.0, 1]	**1.00** [0.4, 2] **4.36** [-0.95, 2]
Votes	0.76 (0.02) 8.42 (0.16)	0.81 (0.00) 7.96 (0.08)	0.82 (0.05) 7.68 (0.03)	0.84 (0.00) 7.87 (0.00)	**0.97** (0.00) 8.06 (0.00)	0.94 [1.0, 2] **7.51** [1.0, 2]	0.94 [0.4, 5] 7.58 [0.3, 2]
XD6	**1.00** (0.00) 6.36 (0.02)	**1.00** (0.00) **6.33** (0.04)	0.74 (0.01) 6.82 (0.01)	**1.00** (0.00) 6.49 (0.01)	**1.00** (0.00) 6.50 (0.00)	0.98 [1.0, 3] 6.55 [1.0, 3]	0.98 [0.0, 3] 6.55 [0.0, 3]
Ave.Acc.	0.673	0.661	0.608	0.742	0.766	0.776	**0.779**

D FORMAL DEFINITIONS AND KNOWN PROPOSITIONS

Definition 1 (χ-deformed m-body approximation)**.** Let $\mathbf{T}$ be a given non-negative normalized tensor and m be a natural number. The (χ, m)-body approximation of $\mathbf{T}$ is defined as

$$\mathbf{P}^{\leq m} \coloneqq \underset{\mathbf{P}\in\mathscr{B}_\chi^{\leq m}}{\arg\min}\, D_\chi(\mathbf{T},\mathbf{P}),$$

where $\mathscr{B}_\chi^{\leq m}$ denotes the set of tensors whose χ-deformed n-body parameters are zero for all $n > m$.

From the definition, the manifold $\mathscr{B}_\chi^{\leq m}$ is e-flat, and the deformed many-body approximation is the m-projection onto this e-flat manifold, which ensures the convexity of the optimization problem. Please refer to Definitions 8 and 9 for the formal definitions of e-flatness and m-projection, respectively. Examples of many-body approximations can be found in Appendix F.1.

Analogously to calling a tensor with reduced rank a low-rank tensor, we call a tensor with reduced interactions a *low-body tensor*.

Definition 2 (Tensor rank (Hitchcock, 1927))**.** The CP-rank, or tensor rank, of D-th order tensor $\mathbf{P}\in\mathbb{R}^{I_1\times\cdots\times I_D}$ is defined as the minimal number R such that $\mathbf{P}$ can be represented exactly as a sum of K rank-1 tensors:

$$\mathrm{rank}_{\mathrm{CP}}(\mathbf{P}) = \min\left\{K \mid \mathbf{P}_{i_1\dots i_D} = \sum_{k=1}^{K} A^{(1)}_{i_1k}A^{(2)}_{i_2k}\dots A^{(D)}_{i_Dk}\right\}.$$

Definition 3 (χ-deformed tensor rank)**.** The χ-deformed tensor rank of the D-th order non-negative tensor $\mathbf{P}\in\mathbb{R}_{\geq 0}^{I_1\times\cdots\times I_D}$ is defined as the minimal number K such that $\mathbf{P}$ can be represented exactly as a sum of K deformed rank-1 tensors:

$$\mathrm{rank}_\chi(\mathbf{P}) = \min\left\{K \mid \mathbf{P}_{i_1\dots i_D} = \sum_{k=1}^{K} A^{(1)}_{i_1k}\circledast A^{(2)}_{i_2k}\circledast\dots\circledast A^{(D)}_{i_Dk}\right\},$$

where $\circledast$ is the appropriately defined χ-deformed product.

It is straightforward to see that the deformed tensor rank for $\chi(x) = x$ is identical to the tensor rank since the corresponding χ-deformed product $\circledast$ becomes the ordinary product. We refer to the approximation of given tensor $\mathbf{T}$ by a (deformed) rank-K tensor $\mathbf{P}$ as a (deformed) rank-K approximation of $\mathbf{T}$.

Although this study focuses on the traditional CP-rank, the deformed Tucker rank (multilinear rank) and the deformed train rank can also be introduced in the same way, namely by replacing the standard product with the deformed product.

D.1 Deformed information geometry and divergences

Definition 4 (Bregman divergence (Bregman, 1967))**.** The Bregman divergence generated by a convex function $\varphi(\boldsymbol{\eta})$ is defined as

$$D_\chi(p,q) = \varphi(\boldsymbol{\eta}^p) - \varphi(\boldsymbol{\eta}^q) - \boldsymbol{\nabla}_{\boldsymbol{\eta}}\varphi(\boldsymbol{\eta}^q)^\top(\boldsymbol{\eta}^p - \boldsymbol{\eta}^q)$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function, and $\boldsymbol{\nabla}_{\boldsymbol{\eta}}\varphi(\boldsymbol{\eta}^q)$ denotes the gradient of φ with respect to $\boldsymbol{\eta}$ at $\boldsymbol{\eta}^q$.

When the convex function $\varphi_\chi(\boldsymbol{\eta})$ is the Shannon entropy, the Bregman divergence recovers the KL-divergence (Shannon, 1948).

Definition 5 (f-divergence (Rényi, 1961))**.** The f-divergence between two discrete probability distributions $\mathbf{P}$ and $\mathbf{Q}$ defined on a finite or countable set Ω is given by

$$D^f(\mathbf{P}, \mathbf{Q}) = \sum_{\boldsymbol{x} \in \Omega} \mathbf{Q}_{\boldsymbol{x}}\, f\left(\frac{\mathbf{P}_{\boldsymbol{x}}}{\mathbf{Q}_{\boldsymbol{x}}}\right), \tag{16}$$

where $f : (0, \infty) \to \mathbb{R}$ is a convex function satisfying $f(1) = 0$.

The total variation distance (Chung et al., 1989), α-divergence (Amari, 2021), Jensen–Shannon divergence, and KL-divergence (Kullback and Leibler, 1951) are known examples of f-divergences.

Definition 6 (Legendre transform (Legendre, 1789))**.** Let $\psi(\boldsymbol{\theta})$ be a convex function on Θ. The Legendre transform $\varphi(\boldsymbol{\eta})$ of $\psi(\boldsymbol{\theta})$ is defined as a convex function such that $\varphi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta} \in \Theta} \left\{\boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta})\right\}$.

The Legendre transform is an involution; that is, the Legendre transform of $\varphi(\boldsymbol{\eta})$ recovers $\psi(\boldsymbol{\theta})$ and vice versa. This relation induces a pair of coordinate systems $\boldsymbol{\eta} = \boldsymbol{\nabla}_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})$ and $\boldsymbol{\theta} = \boldsymbol{\nabla}_{\boldsymbol{\eta}}\varphi(\boldsymbol{\eta})$. The Bregman divergence $D_\psi(p, q)$ is called the dual Bregman divergence of $D_\varphi(p, q)$ when ψ is the Legendre transform of φ.

Definition 7 (χ-divergence)**.** The Bregman divergence generated by the Legendre dual $\varphi_\chi(\boldsymbol{\eta})$ of the χ-free energy $\psi_\chi(\boldsymbol{\theta})$ is referred to as the χ-divergence.

The explicit form of the χ-divergence is given in Equation (8) and Appendix F.3. Strictly speaking, the χ-divergence should be written as $D_{\varphi_\chi}(p, q)$ in this paper; however, for simplicity of notation, we use $D_\chi(p, q)$ throughout the main text.

The Bregman divergence generated by the χ-free energy ψ_χ, which is the normalization factor of the χ-exponential family, is often referred to as the canonical divergence of the χ-deformed exponential family. The χ-divergence is the dual divergence of the canonical divergence of the χ-exponential family $\mathscr{H}_\chi$.

Proposition 1. *Let $\psi(\boldsymbol{\theta})$ be the Legendre transform of the convex function $\varphi(\boldsymbol{\eta})$. Then, it holds that $D_\psi(p, q) = D_\varphi(q, p)$.*

To emphasize the duality between φ and ψ, we also denote $D_\psi(p, q)$ as $D^*_\varphi(q, p)$.

Definition 8 (e- and m-flat manifolds)**.** Within the χ-deformed exponential family $\mathscr{H}_\chi$, a manifold $\mathscr{B}_\chi \subseteq \mathscr{H}_\chi$ defined by a linear constraint on the natural parameter $\boldsymbol{\theta}$ (expectation parameter $\boldsymbol{\eta}$) is referred to as an $e(m)$-flat manifold in $\mathscr{H}_\chi$.

As we show in Proposition 8, the data manifold is a typical example of an m-flat manifold in the ordinary exponential family $\mathscr{H}$.

Definition 9 (e- and m-projections)**.** Let $D_\chi(\cdot, \cdot)$ denote the χ-divergence, and let $\mathscr{B} \subset \mathscr{H}_\chi$. For a point $p \in \mathscr{H}_\chi$, the point $\arg\min_{q \in \mathscr{B}} D_\chi(p, q)$ is called the *m-projection* of p onto $\mathscr{B}$, while $\arg\min_{q \in \mathscr{B}} D_\chi(q, p)$ is called the *e-projection* of p onto $\mathscr{B}$.

The m-projection (resp. e-projection) onto an e-flat (resp. m-flat) manifold $\mathscr{B}$ is a convex optimization problem. Since $\mathscr{B}_\chi^{\leq m}$, defined in Definition 1, is constrained by linear conditions on the natural parameter $\boldsymbol{\theta}$ such that some components of $\boldsymbol{\theta}$ are set to zero, the many-body approximation is a convex optimization problem as it is an m-projection onto this e-flat manifold $\mathscr{B}_\chi^{\leq m}$.

Definition 10 (*em*-algorithm)**.** Given two manifolds $\mathscr{E}_1, \mathscr{E}_2 \subseteq \mathscr{H}_\chi$, the iterative procedure consisting of the following alternating e- and m-projections:

$$p \leftarrow \operatorname*{arg\,min}_{p \in \mathscr{E}_1} D_\chi(q, p),$$
$$q \leftarrow \operatorname*{arg\,min}_{q \in \mathscr{E}_2} D_\chi(p, q),$$

is referred to as the *em*-algorithm between manifolds $\mathscr{E}_1$ and $\mathscr{E}_2$.

Our proposed Algorithm 2 is an *em*-algorithm between the data manifold $\mathscr{D}$ and the low-body manifold $\mathscr{B}_\chi^{\mathrm{CP}}$. Their flatness guarantees the convexity of each projection.

Definition 11 (χ-Fisher information matrix (Amari et al., 2012))**.** Let ψ_χ be a χ-free energy. The χ-Fisher information matrix is defined as

$$\mathbf{G}(\boldsymbol{\theta})_{\boldsymbol{vw}} \coloneqq \frac{\partial}{\partial \theta_{\boldsymbol{v}}} \frac{\partial}{\partial \theta_{\boldsymbol{w}}} \psi_\chi(\boldsymbol{\theta}).$$

D.2 Möbius and Zeta transforms

Definition 12 (Partially ordered set (poset))**.** Let Ω be a set equipped with a binary relation "$\leq$" among members in Ω. The pair $(\Omega, \leq)$ is called a partially ordered set (poset) if the following conditions hold:

1. $\boldsymbol{u} \leq \boldsymbol{u} \Rightarrow \boldsymbol{u} = \boldsymbol{u}$ (reflexivity),
2. $\boldsymbol{u} \leq \boldsymbol{w},\ \boldsymbol{w} \leq \boldsymbol{u} \Rightarrow \boldsymbol{u} = \boldsymbol{w}$ (antisymmetry),
3. $\boldsymbol{u} \leq \boldsymbol{v},\ \boldsymbol{v} \leq \boldsymbol{w} \Rightarrow \boldsymbol{u} \leq \boldsymbol{w}$ (transitivity).

for all $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \Omega$.

The index set Ω of the tensor $\mathbf{P}$ together with the binary relation "$\leq$" introduced in Section 3 forms a poset $(\Omega, \leq)$, which can be immediately verified from the definition. This fact allows us to use the Möbius and zeta functions for coordinate transformations between θ- and η-coordinate systems, as these functions are defined on the direct product space of a poset.

Definition 13 (Möbius and zeta functions (Möbius, 1832))**.** Let Ω be a poset. The zeta function $\zeta : \Omega \times \Omega \to \{0, 1\}$ and the Möbius function $\mu : \Omega \times \Omega \to \mathbb{Z}$ defined as follow.

$$\zeta(x, y) = \begin{cases} 1 & \text{if } x \leq y, \\ 0 & \text{otherwise,} \end{cases} \qquad \mu(x, y) = \begin{cases} 1, & \text{if } x = y, \\ -\sum_{x \leq s < y} \mu(x, s), & \text{if } x < y, \\ 0, & \text{otherwise.} \end{cases}$$

Proposition 2 (Möbius inversion formula (Möbius, 1832))**.** *When functions f, g and h defined on a poset $(\Omega, \leq)$ satisfy*

$$g(x) = \sum_{y \in \Omega} \zeta(y, x) f(y), \quad h(x) = \sum_{y \in \Omega} \zeta(x, y) f(y),$$

the function f can be recovered using the Möbius function as

$$f(x) = \sum_{y \in \Omega} \mu(y, x) g(y), \quad f(x) = \sum_{y \in \Omega} \mu(x, y) h(y).$$

The above Möbius inversion formula is used to derive Equations (11) and (14).

E PROOFS

E.1 Deformed many-body approximation

Proposition 3. *The parameterization in Equation* (2) *forms a χ-exponential family.*

Proof. The χ-exponential family on a sample space Ω is defined by the set of distributions

$$p(\boldsymbol{x}) = \mathrm{Exp}_\chi \left[\boldsymbol{F}(\boldsymbol{x})\boldsymbol{\theta} - \psi_\chi(\boldsymbol{\theta}) \right], \quad x \in \Omega, \tag{17}$$

where $\boldsymbol{F}(\boldsymbol{x}) = (F_1(\boldsymbol{x}), \ldots, F_N(\boldsymbol{x}))$ is appropriately defined functions on Ω and $\psi_\chi(\boldsymbol{\theta})$ is the normalization constant. The real parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)^\top$ are called natural parameters. Let Ω be the set of tensor index

$\Omega = [I_1] \times \cdots \times [I_D]$ and define the tuple of functions $\boldsymbol{F}(\boldsymbol{j}) = (F_{21\ldots1}(\boldsymbol{j}), \ldots, F_{I_1 I_2 \ldots I_D}(\boldsymbol{j}))$ where $F_{\boldsymbol{i}} : \Omega \to \{0, 1\}$ as $F_{\boldsymbol{i}}(\boldsymbol{j}) = \zeta(\boldsymbol{i}, \boldsymbol{j})$ where the zeta function $\zeta(\boldsymbol{i}, \boldsymbol{j})$ is 1 if $\boldsymbol{i} \leq \boldsymbol{j}$ and 0 otherwise. Then the Equation (17) becomes

$$\mathbf{P}_{\boldsymbol{i}} = \mathrm{Exp}_\chi \left[\sum_{\boldsymbol{j} \in \Omega^+}^{N} F_{\boldsymbol{j}}(\boldsymbol{i}) \theta_{\boldsymbol{j}} - \psi_\chi(\boldsymbol{\theta}) \right],$$

which is equivalent to Equation (2). □

Corollary 1. *The set of normalized nonnegative tensors forms a χ-exponential family.*

Proof. It is evident from Proposition 3 and the injective property of the χ-exponential function. □

Since a non-negative normalized tensor is identical to a discrete distribution, the above result is consistent with Theorem 1 in Amari et al. (2012) stating that the discrete distribution forms a χ-exponential family for any positive increasing χ-function. Let $\mathscr{B}_\chi^{\leq m}$ be the set of D-th order (χ,m)-body tensors. Corollary 1 guarantees that $\mathscr{B}_\chi^{\leq D} = \mathscr{H}_\chi$.

Proposition 4. *For the model in Equation* (2)*, its expectation parameter $\boldsymbol{\eta} \coloneqq \partial \psi_\chi(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is given as follows*

$$\boldsymbol{\eta}_{\boldsymbol{v}} = \sum_{\boldsymbol{w} \geq \boldsymbol{v}} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right]. \tag{18}$$

Proof.

$$\begin{aligned}
\frac{\partial}{\partial \theta_{\boldsymbol{v}}} \mathbf{P}_{\boldsymbol{w}} &= \frac{\partial}{\partial \theta_{\boldsymbol{v}}} \mathrm{Exp}_\chi \left[\sum_{\boldsymbol{u} \in \Omega^+} \zeta(\boldsymbol{u}, \boldsymbol{w}) \theta_{\boldsymbol{u}} - \psi_\chi(\boldsymbol{\theta}) \right] \\
&= \left\{ \sum_{\boldsymbol{u} \in \Omega^+} \zeta(\boldsymbol{u}, \boldsymbol{w}) \frac{\partial \theta_{\boldsymbol{w}}}{\partial \theta_{\boldsymbol{v}}} - \frac{\partial \psi_\chi(\boldsymbol{\theta})}{\partial \theta_{\boldsymbol{v}}} \right\} \lambda \left[\sum_{\boldsymbol{u} \in \Omega^+} \zeta(\boldsymbol{u}, \boldsymbol{w}) \theta_{\boldsymbol{u}} - \psi_\chi(\boldsymbol{\theta}) \right] \\
&= \left\{ \zeta(\boldsymbol{v}, \boldsymbol{w}) - \eta_{\boldsymbol{v}} \right\} \chi \left[\mathrm{Exp}_\chi \left[\sum_{\boldsymbol{u} \in \Omega^+} \zeta(\boldsymbol{u}, \boldsymbol{w}) \theta_{\boldsymbol{u}} - \psi_\chi(\boldsymbol{\theta}) \right] \right] \\
&= \left\{ \zeta(\boldsymbol{v}, \boldsymbol{w}) - \eta_{\boldsymbol{v}} \right\} \chi\left[\mathbf{P}_{\boldsymbol{w}}\right],
\end{aligned} \tag{19}$$

where we use the relation $\lambda(t) = \chi\left[\mathrm{Exp}_\chi\left[t\right]\right]$.

By differentiating both sides of the normalization condition $\sum_{\boldsymbol{w} \in \Omega} \mathbf{P}_{\boldsymbol{w}} = 1$ with respect to $\theta_{\boldsymbol{w}}$, we obtain the following.

$$\sum_{\boldsymbol{w} \in \Omega} \left\{ \zeta(\boldsymbol{v}, \boldsymbol{w}) - \eta_{\boldsymbol{v}} \right\} \chi\left[\mathbf{P}_{\boldsymbol{w}}\right] = 0.$$

Then, it holds that

$$\eta_{\boldsymbol{v}} = \frac{\sum_{\boldsymbol{w} \in \Omega} \zeta(\boldsymbol{v}, \boldsymbol{w}) \chi\left[\mathbf{P}_{\boldsymbol{w}}\right]}{\sum_{\boldsymbol{w} \in \Omega} \chi\left[\mathbf{P}_{\boldsymbol{w}}\right]} = \frac{\sum_{\boldsymbol{w} \geq \boldsymbol{v}} \chi\left[\mathbf{P}_{\boldsymbol{w}}\right]}{\sum_{\boldsymbol{w} \in \Omega} \chi\left[\mathbf{P}_{\boldsymbol{w}}\right]} = \sum_{\boldsymbol{w} \geq \boldsymbol{v}} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right].$$

□

Corollary 2. *The χ-escort of the model in Equation* (2) *can be represented by the expectation parameters $\boldsymbol{\eta}$ as follows*

$$\tilde{\chi}\left[\mathbf{P}_{\boldsymbol{v}}\right] = \sum_{\boldsymbol{w} \in \Omega} \mu(\boldsymbol{w}, \boldsymbol{v}) \eta_{\boldsymbol{v}} \tag{20}$$

Proof. This follows from applying the Möbius inversion formula to the relation:

$$\eta_{\boldsymbol{v}} = \sum_{\boldsymbol{w}\in\Omega} \zeta(\boldsymbol{v},\boldsymbol{w})\tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right]. \tag{21}$$

□

Proposition 5. *The χ-entropy $\varphi_\chi(\boldsymbol{\eta})$ is given as*

$$\varphi_\chi(\boldsymbol{\eta}) = \sum_{\boldsymbol{w}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right] \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{w}}\right]. \tag{22}$$

Proof. We consider the divergence generated by the χ-free energy $\psi_\chi(\boldsymbol{\theta})$:

$$\begin{aligned} D^*_\chi(\mathbf{P},\mathbf{Q}) &= \psi_\chi(\boldsymbol{\theta}^{\mathbf{P}}) - \psi_\chi(\boldsymbol{\theta}^{\mathbf{Q}}) - \boldsymbol{\eta}^{\mathbf{Q}}\cdot(\boldsymbol{\theta}^{\mathbf{P}} - \boldsymbol{\theta}^{\mathbf{Q}}) \\ &= \psi_\chi(\boldsymbol{\theta}^{\mathbf{P}}) + \varphi_\chi(\boldsymbol{\eta}^{\mathbf{Q}}) - \boldsymbol{\eta}^{\mathbf{Q}}\cdot\boldsymbol{\theta}^{\mathbf{P}}, \end{aligned} \tag{23}$$

where we use the definition of the Legendre transformation $\varphi_\chi(\boldsymbol{\eta}^{\mathbf{Q}}) = \boldsymbol{\eta}^{\mathbf{Q}}\cdot\boldsymbol{\theta}^{\mathbf{Q}} - \psi_\chi(\boldsymbol{\theta}^{\mathbf{Q}})$ and $\boldsymbol{\eta}^{\mathbf{Q}} = \boldsymbol{\nabla}_{\boldsymbol{\theta}}\psi_\chi(\boldsymbol{\theta}^{\mathbf{Q}})$. Since it holds that $D^*_\chi(\mathbf{P},\mathbf{P}) = 0$, the entropy $\varphi_\chi(\boldsymbol{\eta})$ can be given as

$$\begin{aligned} &\varphi_\chi(\boldsymbol{\eta}) \\ &= \boldsymbol{\theta}\cdot\boldsymbol{\eta} - \psi_\chi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{u}\in\Omega^+} \theta_{\boldsymbol{u}}\eta_{\boldsymbol{u}} - \psi_\chi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{u}\in\Omega^+} \left\{ \sum_{\boldsymbol{v}\in\Omega} \mu(\boldsymbol{v},\boldsymbol{u}) \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{v}}\right] \sum_{\boldsymbol{w}\in\Omega} \zeta(\boldsymbol{u},\boldsymbol{w})\tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right] \right\} - \psi_\chi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{v}\in\Omega}\sum_{\boldsymbol{w}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right] \left\{ \sum_{\boldsymbol{u}\in\Omega^+} \zeta(\boldsymbol{u},\boldsymbol{w})\mu(\boldsymbol{v},\boldsymbol{u}) \right\} \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{v}}\right] - \psi_\chi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{v}\in\Omega}\sum_{\boldsymbol{w}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right] \left\{ \sum_{\boldsymbol{u}\in\Omega} \zeta(\boldsymbol{u},\boldsymbol{w})\mu(\boldsymbol{v},\boldsymbol{u}) - \zeta(\perp,\boldsymbol{w})\mu(\boldsymbol{v},\perp) \right\} \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{v}}\right] - \psi_\chi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{v}\in\Omega}\sum_{\boldsymbol{w}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right] \left\{ \delta_{\boldsymbol{w}\boldsymbol{v}} - \mu(\boldsymbol{v},\perp) \right\} \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{v}}\right] - \psi_\chi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{v}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{v}}\right] \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{v}}\right] - \sum_{\boldsymbol{w}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right] \left\{ \sum_{\boldsymbol{v}\in\Omega} \mu(\boldsymbol{v},\perp) \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{v}}\right] \right\} - \psi_\chi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{v}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{v}}\right] \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{v}}\right] - \sum_{\boldsymbol{w}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right] \theta_\perp - \psi_\chi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{v}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{v}}\right] \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{v}}\right] + \psi_\chi(\boldsymbol{\theta}) - \psi_\chi(\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{v}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{v}}\right] \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{v}}\right], \end{aligned}$$

where we use $\zeta(\perp,\boldsymbol{w}) = 1$ for any $\boldsymbol{w}\in\Omega$ and apply the property $\sum_{\boldsymbol{u}\in\Omega}\zeta(\boldsymbol{u},\boldsymbol{w})\mu(\boldsymbol{v},\boldsymbol{u}) = \delta_{\boldsymbol{v}\boldsymbol{w}}$. □

Proposition 6 (χ-divergence)**.** *The Bregman divergence generated by the χ-entropy $\varphi_\chi(\boldsymbol{\eta})$ is*

$$D_\chi(\mathbf{P},\mathbf{Q}) = \sum_{\boldsymbol{w}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right] \left\{ \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{w}}\right] - \mathrm{Log}_\chi\left[\mathbf{Q}_{\boldsymbol{w}}\right] \right\}. \tag{24}$$

Proof.

$$\begin{aligned}
D_\chi^*(\mathbf{P},\mathbf{Q}) &= \psi_\chi(\boldsymbol{\theta}^{\mathbf{P}}) - \psi_\chi(\boldsymbol{\theta}^{\mathbf{Q}}) + \boldsymbol{\eta}^{\mathbf{Q}} \cdot (\boldsymbol{\theta}^{\mathbf{Q}} - \boldsymbol{\theta}^{\mathbf{P}}) \\
&= \sum_{\boldsymbol{u}\in\Omega} \tilde{\chi}\left[\mathbf{Q}_{\boldsymbol{u}}\right] \left(\psi_\chi(\boldsymbol{\theta}^{\mathbf{P}}) - \psi_\chi(\boldsymbol{\theta}^{\mathbf{Q}})\right) + \sum_{\boldsymbol{x}\in\Omega^+} \sum_{\boldsymbol{u}\in\Omega} \zeta(\boldsymbol{x},\boldsymbol{u}) \tilde{\chi}\left[\mathbf{Q}_{\boldsymbol{u}}\right] \cdot \left(\theta_{\boldsymbol{x}}^{\mathbf{Q}} - \theta_{\boldsymbol{x}}^{\mathbf{P}}\right) \\
&= \sum_{\boldsymbol{u}\in\Omega} \left\{ \sum_{\boldsymbol{x}\in\Omega^+} \zeta(\boldsymbol{x},\boldsymbol{u}) \theta_{\boldsymbol{x}}^{\mathbf{Q}} - \psi_\chi(\boldsymbol{\theta}^{\mathbf{Q}}) - \sum_{\boldsymbol{x}\in\Omega^+} \zeta(\boldsymbol{x},\boldsymbol{u}) \theta_{\boldsymbol{x}}^{\mathbf{P}} + \psi_\chi(\boldsymbol{\theta}^{\mathbf{P}}) \right\} \\
&= \sum_{\boldsymbol{w}\in\Omega} \tilde{\chi}\left[\mathbf{Q}_{\boldsymbol{w}}\right] \left\{ \mathrm{Log}_\chi\left[\mathbf{Q}_{\boldsymbol{w}}\right] - \mathrm{Log}_\chi\left[\mathbf{P}_{\boldsymbol{w}}\right] \right\}.
\end{aligned}$$

Using the relation $D_\chi(\mathbf{P},\mathbf{Q}) = D_\chi^*(\mathbf{Q},\mathbf{P})$, we obtain Equation (24).

□

Theorem 1. *The χ-Fisher information metric*

$$\mathbf{G}(\boldsymbol{\theta})_{\boldsymbol{v}\boldsymbol{w}} \coloneqq \frac{\partial}{\partial\theta_{\boldsymbol{v}}} \frac{\partial}{\partial\theta_{\boldsymbol{w}}} \psi_\chi(\boldsymbol{\theta})$$

is given as

$$\mathbf{G}(\boldsymbol{\theta})_{\boldsymbol{v}\boldsymbol{w}} = \sum_{\boldsymbol{u}\in\Omega} \tilde{\chi}\left[\mathbf{P}_{\boldsymbol{u}}\right] \chi'\left[\mathbf{P}_{\boldsymbol{u}}\right] (\zeta(\boldsymbol{v},\boldsymbol{u}) - \eta_{\boldsymbol{v}})(\zeta(\boldsymbol{w},\boldsymbol{u}) - \eta_{\boldsymbol{w}}) \tag{25}$$

where the function χ' denotes the derivative of the function $\chi(\cdot)$.

Proof. We further differentiate Equation (19) with respect to $\theta_{\boldsymbol{w}}$, leading to the expression below:

$$\begin{aligned}
\frac{\partial}{\partial\theta_{\boldsymbol{w}}} \frac{\partial}{\partial\theta_{\boldsymbol{v}}} \mathbf{P}_{\boldsymbol{u}} &= \frac{\partial}{\partial\theta_{\boldsymbol{w}}} \left\{\zeta(\boldsymbol{v},\boldsymbol{u}) - \eta_{\boldsymbol{v}}\right\} \chi\left[\mathbf{P}_{\boldsymbol{u}}\right] \\
&= -\chi\left[\mathbf{P}_{\boldsymbol{u}}\right] \frac{\partial\eta_{\boldsymbol{v}}}{\partial\theta_{\boldsymbol{w}}} + (\zeta(\boldsymbol{v},\boldsymbol{u}) - \eta_{\boldsymbol{v}}) \frac{\partial\mathbf{P}_{\boldsymbol{u}}}{\partial\theta_{\boldsymbol{w}}} \frac{\partial\chi\left[\mathbf{P}_{\boldsymbol{u}}\right]}{\partial\mathbf{P}_{\boldsymbol{u}}} \\
&= -\frac{\partial\eta_{\boldsymbol{v}}}{\partial\theta_{\boldsymbol{w}}} \chi\left[\mathbf{P}_{\boldsymbol{u}}\right] + (\zeta(\boldsymbol{v},\boldsymbol{u}) - \eta_{\boldsymbol{v}})(\zeta(\boldsymbol{w},\boldsymbol{u}) - \eta_{\boldsymbol{w}}) \chi\left[\mathbf{P}_{\boldsymbol{u}}\right] \chi'\left[\mathbf{P}_{\boldsymbol{u}}\right].
\end{aligned}$$

Because of the normalizing condition $\sum_{\boldsymbol{u}\in\Omega} \mathbf{P}_{\boldsymbol{u}} = 1$, it holds that

$$\sum_{\boldsymbol{u}\in\Omega} \frac{\partial}{\partial\theta_{\boldsymbol{w}}} \frac{\partial}{\partial\theta_{\boldsymbol{v}}} \mathbf{P}_{\boldsymbol{u}} = 0.$$

Thus, we obtain

$$\frac{\partial\eta_{\boldsymbol{v}}}{\partial\theta_{\boldsymbol{w}}} = \frac{1}{\sum_{\boldsymbol{u}\in\Omega} \chi\left[\mathbf{P}_{\boldsymbol{u}}\right]} \sum_{\boldsymbol{u}\in\Omega} \chi\left[\mathbf{P}_{\boldsymbol{u}}\right] \chi'\left[\mathbf{P}_{\boldsymbol{u}}\right] (\zeta(\boldsymbol{v},\boldsymbol{u}) - \eta_{\boldsymbol{v}})(\zeta(\boldsymbol{w},\boldsymbol{u}) - \eta_{\boldsymbol{w}})$$

Using the definitions of the expectation parameter, $\eta_{\boldsymbol{w}} = \partial\psi_\chi(\boldsymbol{\theta})/\partial\theta_{\boldsymbol{w}}$ and the χ-escort distribution $\tilde{\chi}\left[\mathbf{P}_{\boldsymbol{w}}\right] = \chi\left[\mathbf{P}_{\boldsymbol{w}}\right] / \sum_{\boldsymbol{w}\in\Omega} \chi\left[\mathbf{P}_{\boldsymbol{w}}\right]$, we obtain Equation (25). □

This result is a generalization of the known Riemannian metric on the log-linear model on poset (Sugiyama et al., 2017, Theorem 2) since our index set Ω forms a poset.

Corollary 3. *For any $\boldsymbol{v},\boldsymbol{w} \in \Omega$, the first and second derivatives of the χ-divergence $D_\chi(\mathbf{T},\mathbf{P})$ with respect to $\boldsymbol{\theta}^{\mathbf{P}}$ are given as follows.*

$$\frac{\partial}{\partial\theta_{\boldsymbol{v}}^{\mathbf{P}}} D_\chi(\mathbf{T},\mathbf{P}) = \eta_{\boldsymbol{v}}^{\mathbf{P}} - \eta_{\boldsymbol{v}}^{\mathbf{T}}, \tag{26}$$

$$\frac{\partial}{\partial\theta_{\boldsymbol{w}}^{\mathbf{P}}} \frac{\partial}{\partial\theta_{\boldsymbol{v}}^{\mathbf{P}}} D_\chi(\mathbf{T},\mathbf{P}) = \frac{\partial\eta_{\boldsymbol{v}}^{\mathbf{P}}}{\partial\theta_{\boldsymbol{w}}^{\mathbf{P}}}. \tag{27}$$

Proof. Using Equation (23) and Proposition 1, we obtain

$$D_\chi(\mathbf{T},\mathbf{P}) = D^*_\chi(\mathbf{P},\mathbf{T}) = \psi_\chi(\boldsymbol{\theta}^{\mathbf{P}}) + \varphi_\chi(\boldsymbol{\eta}^{\mathbf{T}}) - \boldsymbol{\eta}^{\mathbf{T}} \cdot \boldsymbol{\theta}^{\mathbf{P}}.$$

Thus,

$$\begin{aligned}\frac{\partial}{\partial \theta^{\mathbf{P}}_{\boldsymbol{v}}} D_\chi(\mathbf{T},\mathbf{P}) &= \frac{\partial}{\partial \theta^{\mathbf{P}}_{\boldsymbol{v}}} \psi_\chi(\boldsymbol{\theta}^{\mathbf{P}}) - \frac{\partial}{\partial \theta^{\mathbf{P}}_{\boldsymbol{v}}} \boldsymbol{\eta}^{\mathbf{T}} \cdot \boldsymbol{\theta}^{\mathbf{P}} \\ &= \eta^{\mathbf{P}}_{\boldsymbol{v}} - \eta^{\mathbf{T}}_{\boldsymbol{v}},\end{aligned}$$

and

$$\frac{\partial}{\partial \theta^{\mathbf{P}}_{\boldsymbol{w}}} \frac{\partial}{\partial \theta^{\mathbf{P}}_{\boldsymbol{v}}} D_\chi(\mathbf{T},\mathbf{P}) = \frac{\partial}{\partial \theta^{\mathbf{P}}_{\boldsymbol{w}}} \left(\eta^{\mathbf{P}}_{\boldsymbol{v}} - \eta^{\mathbf{T}}_{\boldsymbol{v}}\right) = \frac{\partial \eta^{\mathbf{P}}_{\boldsymbol{v}}}{\partial \theta^{\mathbf{P}}_{\boldsymbol{w}}}.$$

Since it holds that $\boldsymbol{\eta} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \psi_\chi(\boldsymbol{\theta})$, the second derivative is identical to the χ-Fisher information matrix defined in Definition 11. □

Theorem 2. *The χ-deformed many-body approximation always finds the global optimal solution if the χ-function is differentiable.*

Proof. Recall that both L-BFGS and the Natural gradient method guarantee global optimality if the objective function is convex and its derivative is Lipschitz continuous (Boyd and Vandenberghe, 2004). In the deformed many-body approximation, the model space is restricted by linear constraints in a natural coordinate system because some natural parameters are fixed at zero. This results in a convex optimization problem, as supported by projection theory in information geometry discussed in Section 2.1. The differentiability of the χ-function ensures that the χ-Fisher information matrix (the second derivative of the objective function) is bounded, as evident from Equation (25). Consequently, the first derivative of the objective function in Equation (26) is Lipschitz continuous. Thus, the χ-deformed many-body approximation always finds the global optimal solution. □

E.2 Deformed *em*-algorithm

Proposition 7. *For the Tsallis deformed D-th order low-rank tensor $\mathbf{P}$ with $\chi(s) = s^q$ and $q \to 0$, the rank of the $\mathbf{P}$ is bounded as* $\mathrm{rank}_{\mathrm{CP}}(\mathbf{P}) \leq D$.

Proof. Any deformed low-rank tensor $\mathbf{P} \in \mathbb{R}^{I_1 \times \cdots \times I_D}$ can be expressed as a marginalization of the $(D+1)$-th order tensor $\mathbf{R} \in \mathbb{R}^{I_1 \times \cdots \times I_D \times K}$ over the latent variable k in Equation (10). Since the Tsallis deformed product $x \circledast_q y = (x^{1-q} + y^{1-q} - 1)^{1/(1-q)}$ reduces to the summation $x + y - 1$ as $q \to 0$, the tensor $\mathbf{R}$ in Equation (10) becomes

$$\begin{aligned}\mathbf{R}_{i_1 \dots i_D k} &= \frac{1}{Z_\chi} \circledast_q X^{(1)}_{i_1 k} \circledast_q X^{(2)}_{i_2 k} \circledast_q \dots \circledast_q X^{(D)}_{i_D k} \\ &= \frac{1}{Z_\chi} + X^{(1)}_{i_1 k} + X^{(2)}_{i_2 k} + \cdots + X^{(D)}_{i_D k} - D.\end{aligned}$$

Marginalizing over the latent variable k gives the deformed low-rank tensor as

$$\begin{aligned}\mathbf{P}_{i_1 \dots i_D} &= \sum_{k=1}^{K} \mathbf{R}_{i_1 \dots i_D k} \\ &= \sum_{k=1}^{K} \left(\frac{1}{Z_\chi} + X^{(1)}_{i_1 k} + X^{(2)}_{i_1 k} + \cdots + X^{(D)}_{i_1 k} - D \right) \\ &= u^{(1)}_{i_1} + u^{(2)}_{i_2} + \cdots + u^{(D)}_{i_D},\end{aligned}$$

where each element of the vector $\boldsymbol{u}^{(d)} \in \mathbb{R}^{I_d}$ is defined as

$$u^{(d)}_{i_d} = \frac{1}{K} \left(\frac{1}{Z_\chi} - D \right) + \sum_{k=1}^{K} X^{(d)}_{i_d k}.$$

Hence, we can write $\mathbf{R}$ as

$$\mathbf{R} = \boldsymbol{u}^{(1)} \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} + \mathbf{1} \otimes \boldsymbol{u}^{(2)} \otimes \cdots \otimes \mathbf{1} + \cdots + \mathbf{1} \otimes \mathbf{1} \otimes \cdots \otimes \boldsymbol{u}^{(D)},$$

where the symbol $\otimes$ denotes the standard outer product. Since each term is a rank-1 tensor, we conclude that $\mathrm{rank}_{\mathrm{CP}}(\mathbf{P}) \leq D$. □

Corollary 4. *For $\chi(s) = s^q$, the deformed rank of any χ-deformed low-rank tensor $\mathbf{P}$ is 1 in the limit of $q \to 0$.*

Proof. In the limit of $q \to 0$, any tensor whose deformed rank is at most K can be written as

$$\begin{aligned}\mathbf{P}_{i_1 \ldots i_D} &= \sum_{k=1}^{K} \left(\frac{1}{Z_\chi} + X^{(1)}_{i_1 k} + \cdots + X^{(D)}_{i_D k} - D \right) \\ &= \sum_{k=1}^{1} \left(\frac{1}{Z_\chi} + Y^{(1)}_{i_1 k} + \cdots + Y^{(D)}_{i_D k} - D \right),\end{aligned}$$

where we define $Y^{(d)}_{i_d 1} = \sum_{k=1}^{K} X^{(d)}_{i_d k}$. Thus, it holds that $\mathrm{rank}_\chi(\mathbf{P}) = 1$ for any $\mathbf{P}$ in the limit $q \to 0$. □

This result suggests that in the Tsallis deformed low-rank approximation, the parameter q controls the regularization of the deformed low-rank structure, while the deformed rank K controls the number of model parameters. Appropriately choosing these parameters leads to stable learning, as demonstrated in Section 5. To our knowledge, using the deformed product in tensor factorization to achieve regularization has not been explored before.

Proposition 8. *For a given tensor $\mathbf{T}$, the data manifold $\mathscr{D} = \left\{ \mathbf{Q} \mid \sum_k \mathbf{Q}_{i_1 \ldots i_D k} = \mathbf{T}_{i_1 \ldots i_D} \right\}$ is m-flat in the coordinate system of the exponential family $\mathscr{H}$.*

Proof. For the sub-index set $\Omega' = [I_1] \times \cdots \times [I_D] \times [1] \subset \Omega$, the constraint defining the data manifold, $\sum_k \mathbf{Q}_{i_1 \ldots i_D k} = \mathbf{T}_{i_1 \ldots i_D}$, can be written as

$$\sum_{\boldsymbol{w} \in \Omega'} \mu(\boldsymbol{i}, \boldsymbol{w}) \boldsymbol{\eta}^{\mathbf{Q}}_{\boldsymbol{w}} = \mathbf{T}_{\boldsymbol{i}}, \tag{28}$$

where $\boldsymbol{\eta}^{\mathbf{Q}}$ denotes the expectation parameter of the exponential family, defined by

$$\boldsymbol{\eta}^{\mathbf{Q}}_{\boldsymbol{w}} = \sum_{\boldsymbol{j} \geq \boldsymbol{w}} \mathbf{Q}_{\boldsymbol{j}}.$$

Since Equation (28) is linear in $\boldsymbol{\eta}$, the data manifold $\mathscr{D}$ is m-flat in the coordinate system of the exponential family $\mathscr{H}$. □

Proposition 9 (Monotonicity property of the f-divergence)**.** *For given $\mathbf{T} \in \mathscr{S}(D)$ and any $\mathbf{Q} \in \mathscr{D} = \left\{ \mathbf{Q} \mid \sum_k \mathbf{Q}_{i_1 \ldots i_D k} = \mathbf{T}_{i_1 \ldots i_D} \right\}$, it holds that*

$$D^f(\mathbf{T}, \mathbf{P}) \leq D^f(\mathbf{Q}, \mathbf{R})$$

where $\mathbf{P} \in \mathscr{S}(D)$ satisfies $\mathbf{P}_{\boldsymbol{i}} = \sum_k \mathbf{R}_{i_1 \ldots i_D k}$.

Proof.

$$\begin{aligned}
D^f(\mathbf{Q},\mathbf{R}) &= \sum_{\boldsymbol{i}\in\Omega}\sum_k \mathbf{R}_{i_1\dots i_D k} f\left(\frac{\mathbf{Q}_{i_1\dots i_D k}}{\mathbf{R}_{i_1\dots i_D k}}\right) \\
&= \sum_{\boldsymbol{i}\in\Omega}\sum_k \mathbf{P}_{i_1\dots i_D}\frac{\mathbf{R}_{i_1\dots i_D k}}{\mathbf{P}_{i_1\dots i_D}} f\left(\frac{\mathbf{Q}_{i_1\dots i_D k}}{\mathbf{R}_{i_1\dots i_D k}}\right) \\
&= \sum_{\boldsymbol{i}\in\Omega}\mathbf{P}_{i_1\dots i_D}\sum_k \frac{\mathbf{R}_{i_1\dots i_D k}}{\mathbf{P}_{i_1\dots i_D}} f\left(\frac{\mathbf{Q}_{i_1\dots i_D k}}{\mathbf{R}_{i_1\dots i_D k}}\right) \\
&\geq \sum_{\boldsymbol{i}\in\Omega}\mathbf{P}_{i_1\dots i_D} f\left(\sum_k \frac{\mathbf{R}_{i_1\dots i_D k}}{\mathbf{P}_{i_1\dots i_D}}\frac{\mathbf{Q}_{i_1\dots i_D k}}{\mathbf{R}_{i_1\dots i_D k}}\right) \\
&= \sum_{\boldsymbol{i}\in\Omega}\mathbf{P}_{i_1\dots i_D} f\left(\frac{\sum_k \mathbf{Q}_{i_1\dots i_D k}}{\mathbf{P}_{i_1\dots i_D}}\right) \\
&= \sum_{\boldsymbol{i}\in\Omega}\mathbf{P}_{i_1\dots i_D} f\left(\frac{\mathbf{T}_{i_1\dots i_D}}{\mathbf{P}_{i_1\dots i_D}}\right) = D^f(\mathbf{T},\mathbf{P})
\end{aligned}$$

where the following relation, as defined by Jensen's inequality (Jensen, 1906), is used:

$$f\left(\sum_{m=1}^M \lambda_m x_m\right) \leq \sum_{m=1}^M \lambda_m f(x_m), \tag{29}$$

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

Proposition 10. *For a given non-negative normalized tensor $\mathbf{T}\in\mathscr{S}(D)$, let $\mathscr{D}$ be the data manifold $\mathscr{D} = \left\{\mathbf{Q} \mid \sum_k \mathbf{Q}_{i_1\dots i_D,k} = \mathbf{T}_{i_1\dots i_D}\right\}$. For any f-divergence and tensor $\mathbf{R}\in\mathscr{S}(D+1)$, the global optimal tensor $\mathbf{Q}^*\in\mathscr{S}(D+1)$ satisfying $\mathbf{Q}^* = \arg\min_{\mathbf{Q}\in\mathscr{D}} D^f(\mathbf{Q},\mathbf{R})$ is given as follows:*

$$\mathbf{Q}^*_{i_1\dots i_D} = \frac{\mathbf{T}_{\boldsymbol{i}}\mathbf{R}_{i_1\dots i_D k}}{\sum_k \mathbf{R}_{i_1\dots i_D k}}. \tag{30}$$

Proof.

$$\begin{aligned}
D^f(\mathbf{Q}^*,\mathbf{R}) &= \sum_{\boldsymbol{i}\in\Omega}\sum_k \mathbf{R}_{i_1\dots i_D k} f\left(\frac{\mathbf{Q}^*_{i_1\dots i_D k}}{\mathbf{R}_{i_1\dots i_D k}}\right) \\
&= \sum_{\boldsymbol{i}\in\Omega}\sum_k \mathbf{R}_{i_1\dots i_D k} f\left(\frac{\mathbf{T}_{i_1\dots i_D}}{\sum_k \mathbf{R}_{i_1\dots i_D k}}\right) \\
&= \sum_{\boldsymbol{i}\in\Omega}\sum_k \mathbf{R}_{i_1\dots i_D k} f\left(\frac{\mathbf{T}_{i_1\dots i_D}}{\mathbf{P}_{i_1\dots i_D}}\right) \\
&= \sum_{\boldsymbol{i}\in\Omega}\mathbf{P}_{i_1\dots i_D} f\left(\frac{\mathbf{T}_{i_1\dots i_D}}{\mathbf{P}_{i_1\dots i_D}}\right) \\
&= D^f(\mathbf{T},\mathbf{P}),
\end{aligned}$$

where we define the tensor $\mathbf{P}\in\mathscr{S}(D)$ as $\mathbf{P}_{i_1\dots i_D} = \sum_k \mathbf{R}_{i_1\dots i_D k}$. □

To the best of our knowledge, Zhang et al. (2019) first demonstrated that the bound of the f-divergence can be optimized in closed form and the solution makes the inequality tight. Since their analysis focuses on variational inference, we restate the proposition here to ensure consistency in the context of tensor decomposition.

Theorem 3. *The f-divergence-based em-algorithm for deformed low-rank approximation always converges.*

Proof. For a given tensor $\mathbf{T}$, the algorithm optimizes the f-divergence $D^f(\mathbf{T}, \mathbf{P})$ over the deformed low-rank tensor $\mathbf{P}$ by iteratively minimizing the upper bound $D^f(\mathbf{Q}, \mathbf{R})$, as stated in Proposition 9.

In the e-step of iteration t, we minimize the upper bound $D^f(\mathbf{Q}^{t-1}, \mathbf{R}^{t-1})$ over the tensor $\mathbf{Q}^{t-1}$ and then update $\mathbf{Q}^{t-1}$ by the optimal tensor $\mathbf{Q}^t \in \mathscr{D}$. Thus,

$$D^f(\mathbf{Q}^t, \mathbf{R}^{t-1}) \leq D^f(\mathbf{Q}^{t-1}, \mathbf{R}^{t-1}). \tag{31}$$

In the subsequent m-step, we minimize the current upper bound $D^f(\mathbf{Q}^t, \mathbf{R}^{t-1})$ over the low-body tensor $\mathbf{R}$.

$$D^f(\mathbf{Q}^t, \mathbf{R}^t) \leq D^f(\mathbf{Q}^t, \mathbf{R}^{t-1}). \tag{32}$$

In the e-step of the next iteration, we optimize the updated bound $D^f(\mathbf{Q}^t, \mathbf{R}^t)$ over $\mathbf{Q}$, giving

$$D^f(\mathbf{Q}^{t+1}, \mathbf{R}^t) \leq D^f(\mathbf{Q}^t, \mathbf{R}^t). \tag{33}$$

Since the e-step tightens the upper bound to the original objective function, as shown in Proposition 10, it holds that $D^f(\mathbf{Q}^{t+1}, \mathbf{R}^t) = D^f(\mathbf{T}, \mathbf{P}^t)$ and $D^f(\mathbf{Q}^t, \mathbf{R}^{t-1}) = D^f(\mathbf{T}, \mathbf{P}^{t-1})$, where $\mathbf{P}^t$ is obtained by marginalizing the low-body tensor $\mathbf{R}^t$ over the index k, i.e., $\mathbf{P}^t_{\boldsymbol{i}} = \sum_k \mathbf{R}^t_{\boldsymbol{i}k}$. Finally, we obtain

$$D^f(\mathbf{T}, \mathbf{P}^t) \leq D^f(\mathbf{Q}^t, \mathbf{R}^t) \leq D^f(\mathbf{T}, \mathbf{P}^{t-1}). \tag{34}$$

Thus, the algorithm converges because the f-divergence objective is bounded below and decreases monotonically at each iteration. □

E.3 Perturbation analysis

The following statement is used for the perturbation analysis in Appendix G.1.

Lemma 1. *We assume the function $L_\chi(\epsilon) \coloneqq D_\chi(\mathbf{Q}^\epsilon, \mathbf{P})$ is three times differentiable on $\mathbb{R}$ and consider Gaussian distributed perturbations $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and thus $\sigma = O_p(\epsilon)$ with some arbitrary small $\sigma \geq 0$. For*

$$\mathbf{P}_{ij} = \frac{1}{Z^{\mathbf{P}}_\chi} \circledast_\chi u_i \circledast_\chi v_j, \quad \mathbf{Q}^\epsilon_{ij} = \frac{1}{Z^{\mathbf{Q}^\epsilon}_\chi} \circledast_\chi u_i \circledast_\chi \left(v_j + \epsilon\delta v_j\right),$$

it holds almost surely that

$$\mathbb{E}_\epsilon[L_\chi(\epsilon)] = D_\chi(\mathbf{Q}^{\epsilon=0}, \mathbf{P}) + \sigma^2 L''_\chi(0) + O_p(\sigma^3), \tag{35}$$

where $L''_\chi(\epsilon)$ denotes second derivative of the function $L_\chi(\epsilon)$.

Proof. Let I be the finite interval around the origin. Since we assume the function $L_\chi(\epsilon)$ is three times differentiable, there is a non-negative constant $C \geq 0$ that satisfies

$$0 \leq |L'_\chi(\epsilon)| \leq C,$$

for any $\epsilon \in I$. Thus, we have $-\epsilon C \leq \epsilon L'_\chi(0) \leq \epsilon C$ for $\epsilon \in I$. Taking expectation with respect to ϵ and applying Chebyshev inequality (Bienaymé, 1853), we obtain $\mathbb{E}_\epsilon[L'_\chi(0)] \to 0$, and the statement follows by Taylor expansion of $L_\chi(\epsilon)$ up to third order around the mean value of ϵ, i.e., $\epsilon = 0$. □

F EXAMPLES

For the sake of clarity and readability, we present here a few illustrative examples that correspond to special cases of the general theory discussed in the main text.

F.1 Deformed many-body approximation for fourth-order tensor

We here provide an example of the (χ, 1)-body and the (χ, 2)-body approximations for $D = 4$. The model $\mathbf{P} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4}$ on the index set $\Omega = [I_1] \times [I_2] \times [I_3] \times [I_4]$ in Equation (2) can be written as

$$\mathbf{P}_{i_1 i_2 i_3 i_4} = \mathrm{Exp}_\chi \left[-E_{i_1 i_2 i_3 i_4} - \psi_\chi(\boldsymbol{\theta}) \right],$$

where the energy function can be given as

$$\begin{aligned} E_{i_1 i_2 i_3 i_4} = & \underbrace{E^{(1)}_{i_1} + E^{(2)}_{i_2} + E^{(3)}_{i_3} + E^{(4)}_{i_4}}_{\text{one-body}} \\ & + \underbrace{E^{(12)}_{i_1 i_2} + E^{(13)}_{i_1 i_3} + E^{(14)}_{i_1 i_4} + E^{(23)}_{i_2 i_3} + E^{(24)}_{i_2 i_4} + E^{(34)}_{i_3 i_4}}_{\text{two-body}} \\ & + \underbrace{E^{(123)}_{i_1 i_2 i_3} + E^{(124)}_{i_1 i_2 i_4} + E^{(134)}_{i_1 i_3 i_4} + E^{(234)}_{i_2 i_3 i_4}}_{\text{three-body}} \\ & + \underbrace{E^{(1234)}_{i_1 i_2 i_3 i_4}}_{\text{four-body}}. \end{aligned}$$

An m-body term governs the interactions among m modes. For example, $E^{(134)}$ governs the third-order (three-body) interaction among first, third, and fourth modes. Each term can be represented as a sum over the corresponding m-body parameters as follows.

One-body energy:

$$E^{(1)}_{i_1} = \sum_{i_1=2}^{I_1} \theta_{i_1 111}, \quad E^{(2)}_{i_2} = \sum_{i_2=2}^{I_2} \theta_{1 i_2 11}, \quad E^{(3)}_{i_3} = \sum_{i_3=2}^{I_3} \theta_{111 i_3 1}, \quad E^{(4)}_{i_4} = \sum_{i_4=2}^{I_4} \theta_{111 i_4},$$

Two-body energy:

$$\begin{aligned} E^{(12)}_{i_1 i_2} &= \sum_{i_1=2}^{I_1} \sum_{i_2=2}^{I_2} \theta_{i_1 i_2 11}, \quad E^{(13)}_{i_1 i_3} = \sum_{i_1=2}^{I_1} \sum_{i_3=2}^{I_3} \theta_{i_1 1 i_3 1}, \quad E^{(14)}_{i_1 i_4} = \sum_{i_1=2}^{I_1} \sum_{i_4=2}^{I_4} \theta_{i_1 11 i_4}, \\ E^{(23)}_{i_2 i_3} &= \sum_{i_2=2}^{I_2} \sum_{i_3=2}^{I_3} \theta_{1 i_2 i_3 1}, \quad E^{(24)}_{i_2 i_4} = \sum_{i_2=2}^{I_2} \sum_{i_4=2}^{I_4} \theta_{1 i_2 1 i_4}, \quad E^{(34)}_{i_3 i_4} = \sum_{i_3=2}^{I_3} \sum_{i_4=2}^{I_4} \theta_{11 i_3 i_4}, \end{aligned}$$

Three-body energy:

$$\begin{aligned} E^{(123)}_{i_1 i_2 i_3} &= \sum_{i_1=2}^{I_1} \sum_{i_2=2}^{I_2} \sum_{i_3=2}^{I_3} \theta_{i_1 i_2 i_3 1}, \quad E^{(124)}_{i_1 i_2 i_4} = \sum_{i_1=2}^{I_1} \sum_{i_2=2}^{I_2} \sum_{i_4=2}^{I_4} \theta_{i_1 i_2 1 i_4}, \\ E^{(134)}_{i_1 i_3 i_4} &= \sum_{i_1=2}^{I_1} \sum_{i_3=2}^{I_3} \sum_{i_4=2}^{I_4} \theta_{i_1 1 i_3 i_4}, \quad E^{(234)}_{i_2 i_3 i_4} = \sum_{i_2=2}^{I_2} \sum_{i_3=2}^{I_3} \sum_{i_4=2}^{I_4} \theta_{1 i_2 i_3 i_4}, \end{aligned}$$

Four-body energy:

$$E^{(1234)}_{i_1 i_2 i_3 i_4} = \sum_{i_1=2}^{I_1} \sum_{i_2=2}^{I_2} \sum_{i_3=2}^{I_3} \sum_{i_4=2}^{I_4} \theta_{i_1 i_2 i_3 i_4},$$

The m-body approximation approximates the tensor using only the energy terms of m-th order (m-body) or lower. By imposing the constraint that the $n(> m)$-body parameters are zero, higher-order interactions are removed, as shown below.

F.1.1 Deformed one-body approximation

For example, in the one-body approximation, the model manifold is given as follows

$$\mathscr{B}_{\chi}^{\leq 1} = \left\{ \mathbf{P} \mid \theta_{\boldsymbol{v}}^{\mathbf{P}} = 0 \text{ for all } \boldsymbol{v} \notin \Omega_{B_1} \right\},$$

where the sub index set $\Omega_{B_1} \subset \Omega$ is defined such that $(\theta_{\boldsymbol{v}})_{\boldsymbol{v}\in\Omega_{B_1}}$ becomes the set of all deformed one-body parameters, that is

$$\begin{aligned}\Omega_{B_1} = &\left([2:I_1]\times[1]\times[1]\times[1]\right)\\ &\cup\left([1]\times[2:I_2]\times[1]\times[1]\right)\\ &\cup\left([1]\times[1]\times[2:I_3]\times[1]\right)\\ &\cup\left([1]\times[1]\times[1]\times[2:I_4]\right),\end{aligned}$$

and $[A:B]$ denotes the set of integers greater than $A \in \mathbb{N}$ and less than $B \in \mathbb{N}$, i.e., $[A:B] = \{A, A+1, A+2, \ldots, B\}$. It is straightforward to see that $(\theta_{\boldsymbol{v}})_{\boldsymbol{v}\in\Omega_{B_1}}$ becomes the set of all one-body parameters. Thus, for any tensor $\mathbf{P}^{\leq 1}$ in the model manifold $\mathscr{B}_{\chi}^{\leq 1}$, all two-, three-, and four-body energies become 0 and they can be written as

$$\begin{aligned}\mathbf{P}^{\leq 1}_{i_1 i_2 i_3 i_4} &= \mathrm{Exp}_{\chi}\left[E^{(1)}_{i_1} + E^{(2)}_{i_2} + E^{(3)}_{i_3} + E^{(4)}_{i_4} - \psi_{\chi}(\boldsymbol{\theta})\right]\\ &= \mathrm{Exp}_{\chi}\left[E^{(1)}_{i_1}\right] \circledast \mathrm{Exp}_{\chi}\left[E^{(2)}_{i_2}\right] \circledast \mathrm{Exp}_{\chi}\left[E^{(3)}_{i_3}\right] \circledast \mathrm{Exp}_{\chi}\left[E^{(4)}_{i_4}\right] \circledast \mathrm{Exp}_{\chi}\left[-\psi_{\chi}(\boldsymbol{\theta})\right]\\ &= \frac{1}{Z_{\chi}} \circledast X^{(1)}_{i_1} \circledast X^{(2)}_{i_2} \circledast X^{(3)}_{i_3} \circledast X^{(4)}_{i_4}, \end{aligned} \tag{36}$$

where we used the deformed algebra $\mathrm{Exp}_{\chi}[A+B] = \mathrm{Exp}_{\chi}[A] \circledast \mathrm{Exp}_{\chi}[B]$ and define the normalizer Z_{χ} as $Z_{\chi}^{-1} = \mathrm{Exp}_{\chi}\left[-\psi(\boldsymbol{\theta})\right]$. Each factor is given as $X^{(d)}_{i_d} = \mathrm{Exp}_{\chi}\left[E^{(d)}_{i_d}\right]$. Each tensor in the manifold $\mathscr{B}_{\chi}^{\leq 1}$ is called a $(\chi, 1)$-body tensor. When $\chi(s) = s$, the deformed product $\circledast$ is replaced by ordinary product and it hold that $\mathrm{rank}_{\mathrm{CP}}(\mathbf{P}) = 1$ since the tensor $\mathbf{P}^{\leq 1} \in \mathscr{B}_{\chi}^{\leq 1}$ in Equation (36) recovers

$$\mathbf{P}^{\leq 1} = \tilde{X}^{(1)} \otimes \tilde{X}^{(2)} \otimes \tilde{X}^{(3)} \otimes \tilde{X}^{(4)},$$

for $\tilde{X}^{(d)}_{i_d} = X^{(d)}_{i_d} / \sqrt[4]{Z}$ and the standard tensor product $\otimes$. In this sense, the deformed many-body approximation can be seen as an extension of the standard rank-1 approximation.

F.1.2 Deformed two-body approximation

In the same way, the model manifold for the two-body approximation is given as

$$\mathscr{B}_{\chi}^{\leq 2} = \left\{ \mathbf{P} \mid \theta_{\boldsymbol{v}}^{\mathbf{P}} = 0, \text{ for all } \boldsymbol{v} \notin \Omega_{B_2} \right\},$$

where the sub index set $\Omega_{B_2} \subset \Omega$ is defined as follows:

$$\begin{aligned}\Omega_{B_2} = &\Omega_{B_1}\\ &\cup\left([2:I_1]\times[2:1_2]\times[1]\times[1]\right)\\ &\cup\left([2:I_1]\times[1]\times[2:I_3]\times[1]\right)\\ &\cup\left([2:I_1]\times[1]\times[1]\times[2:I_4]\right)\\ &\cup\left([1]\times[2:I_2]\times[2:I_3]\times[1]\right)\\ &\cup\left([1]\times[2:I_2]\times[1]\times[2:I_4]\right)\\ &\cup\left([1]\times[1]\times[2:I_3]\times[2:I_4]\right).\end{aligned}$$

It is straightforward to see that $(\theta_{\boldsymbol{v}})_{\boldsymbol{v}\in\Omega_{B_2}}$ becomes the set of all one-body and two-body parameters. Thus, both of all three-, and four-body energies become 0 and the tensor $\mathbf{P}^{\leq 2}$ in the model $\mathscr{B}_{\chi}^{\leq 2}$ can be written as

$$\begin{aligned}
\mathbf{P}^{\leq 2}_{i_1i_2i_3i_4} &= \mathrm{Exp}_\chi\left[E^{(1)}_{i_1}+E^{(2)}_{i_2}+E^{(3)}_{i_3}+E^{(4)}_{i_4}+E^{(12)}_{i_1i_2}+E^{(13)}_{i_1i_3}+E^{(14)}_{i_1i_4}+E^{(23)}_{i_2i_3}+E^{(24)}_{i_2i_4}+E^{(34)}_{i_3i_4}-\psi_\chi(\boldsymbol{\theta})\right]\\
&= \mathrm{Exp}_\chi\left[\frac{1}{3}E^{(1)}_{i_1}+E^{(1,2)}_{i_1i_2}+\frac{1}{3}E^{(2)}_{i_2}\right]\circledast\mathrm{Exp}_\chi\left[\frac{1}{3}E^{(1)}_{i_1}+E^{(1,3)}_{i_1i_3}+\frac{1}{3}E^{(3)}_{i_3}\right]\\
&\circledast\mathrm{Exp}_\chi\left[\frac{1}{3}E^{(1)}_{i_1}+E^{(1,4)}_{i_1i_4}+\frac{1}{3}E^{(4)}_{i_4}\right]\circledast\mathrm{Exp}_\chi\left[\frac{1}{3}E^{(2)}_{i_2}+E^{(2,3)}_{i_2i_3}+\frac{1}{3}E^{(3)}_{i_3}\right]\\
&\circledast\mathrm{Exp}_\chi\left[\frac{1}{3}E^{(2)}_{i_2}+E^{(2,4)}_{i_2i_4}+\frac{1}{3}E^{(4)}_{i_4}\right]\circledast\mathrm{Exp}_\chi\left[\frac{1}{3}E^{(3)}_{i_3}+E^{(3,4)}_{i_3i_4}+\frac{1}{3}E^{(4)}_{i_4}\right]\circledast\mathrm{Exp}_\chi\left[-\psi_\chi(\boldsymbol{\theta})\right]\\
&= \frac{1}{Z_\chi}\circledast X^{(1,2)}_{i_1i_2}\circledast X^{(1,3)}_{i_1i_3}\circledast X^{(1,4)}_{i_1i_4}\circledast X^{(2,3)}_{i_2i_3}\circledast X^{(2,4)}_{i_2i_4}\circledast X^{(3,4)}_{i_3i_4l},
\end{aligned}$$

where one- and two-body factors are given as

$$X^{(d)}_{i_d}=\mathrm{Exp}_\chi\left[E^{(d)}_{i_d}\right],\quad X^{(k,l)}_{i_ki_l}=\mathrm{Exp}_\chi\left[\frac{1}{3}E^{(k)}_{i_k}+E^{(k,l)}_{i_ki_l}+\frac{1}{3}E^{(l)}_{i_l}\right],$$

respectively, and the normalizer Z_χ is defined as $1/Z_\chi=\mathrm{Exp}_\chi\left[-\psi_\chi(\boldsymbol{\theta})\right]$. Each tensor in the manifold $\mathscr{B}_{\chi}^{\leq 2}$ is called a $(\chi,2)$-body tensor. We note that, even if the deformed product is replaced with the ordinary product, the two-body tensor is not necessarily of low-rank.

As is clear from the definition, for the set of m-body tensors $\mathscr{B}_{\chi}^{\leq m}$ it holds that $\mathscr{B}_{\chi}^{\leq m}\subseteq\mathscr{B}_{\chi}^{\leq m+1}$. From the extended Pythagorean theorem (Amari, 2021), it immediately follows that the projection onto $\mathscr{B}^{\leq m}$ does not necessarily coincide with the composition of the projections onto $\mathscr{B}^{\leq m+1}$ and then onto $\mathscr{B}^{\leq m}$.

F.2 Parameter transformation

In the proposed Algorithm 1, we need to convert the tensor $\mathbf{P}\in\mathbb{R}^{I_1\times\cdots\times I_D}$ into its natural parameters $\boldsymbol{\theta}$ using Equation (12). We here show an example of the transformation with $D=2,3$ based on Equation (12).

$$\begin{aligned}
\theta_{i_1,i_2} =\ & \mathrm{Log}_\chi\left[\mathbf{P}_{i_1,i_2}\right]-\mathrm{Log}_\chi\left[\mathbf{P}_{i_1-1,i_2}\right]-\mathrm{Log}_\chi\left[\mathbf{P}_{i_1,i_2-1}\right]+\mathrm{Log}_\chi\left[\mathbf{P}_{i_1-1,i_2-1}\right],\\
\theta_{i_1,i_2,i_3} =\ & \mathrm{Log}_\chi\left[\mathbf{P}_{i_1,i_2,i_3}\right]-\mathrm{Log}_\chi\left[\mathbf{P}_{i_1-1,i_2,i_3}\right]-\mathrm{Log}_\chi\left[\mathbf{P}_{i_1,i_2-1,i_3}\right]-\mathrm{Log}_\chi\left[\mathbf{P}_{i_1,i_2,i_3-1}\right]\\
&+\mathrm{Log}_\chi\left[\mathbf{P}_{i_1-1,i_2-1,i_3}\right]+\mathrm{Log}_\chi\left[\mathbf{P}_{i_1,i_2-1,i_3-1}\right]+\mathrm{Log}_\chi\left[\mathbf{P}_{i_1,i_2-1,i_3-1}\right]-\mathrm{Log}_\chi\left[\mathbf{P}_{i_1-1,i_2-1,i_3-1}\right],
\end{aligned}$$

where we assume $\mathbf{P}_{0,i_2}=\mathbf{P}_{i_1,0}=1$ and $\mathbf{P}_{i_1,i_2,0}=\mathbf{P}_{i_1,0,i_3}=\mathbf{P}_{0,i_2,i_3}=1$. We can also obtain expectation parameters $\boldsymbol{\eta}$ using Equation (14) as follows:

$$\begin{aligned}
\tilde{\chi}\left[\mathbf{P}_{i_1,i_2}\right] &= \eta_{i_1,i_2}-\eta_{i_1+1,i_2}-\eta_{i_1,i_2+1}+\eta_{i_1+1,i_2+1},\\
\tilde{\chi}\left[\mathbf{P}_{i_1,i_2,i_3}\right] &= \eta_{i_1,i_2,i_3}-\eta_{i_1+1,i_2,i_3}-\eta_{i_1,i_2+1,i_3}-\eta_{i_1,i_2,i_3+1}\\
&\quad+\eta_{i_1+1,i_2+1,i_3}+\eta_{i_1+1,i_2,i_3+1}+\eta_{i_1,i_2+1,i_3+1}-\eta_{i_1+1,i_2+1,i_3+1},
\end{aligned}$$

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

These transformations can be efficiently carried out with `cumsum` function in NumPy (Harris et al., 2020) and are not a computational bottleneck.

F.3 χ-divergence

When $\chi(x)=x^q$, the χ-divergence in Equation (8) is called the Tsallis divergence and it can be given as

$$D_q(\mathbf{T},\mathbf{P})=\frac{1}{h^q(\mathbf{T})}\frac{1}{1-q}\sum_{\boldsymbol{u}\in\Omega}\left\{1-\mathbf{T}^q_{\boldsymbol{u}}\mathbf{P}^{1-q}_{\boldsymbol{u}}\right\},\tag{37}$$

where $h_q(\mathbf{T})=\sum_{\boldsymbol{u}\in\Omega}\mathbf{T}^q_{\boldsymbol{u}}$. The Tsallis divergence becomes the KL divergence for $q\to 1$. As discussed in Section 3.1, its optimization is equivalent to that of Amari's α-divergence and Rényi-α divergence (Van Erven and Harremos, 2014).

When $\chi(x) = x/\cosh(\kappa \log x)$, the χ-divergence in Equation (8) can be written as

$$D_\kappa(\mathbf{T}, \mathbf{P}) = \frac{1}{\kappa \sum_{\boldsymbol{u}\in\Omega} \left(\frac{\mathbf{P}_{\boldsymbol{u}}}{\cosh(\kappa \log \mathbf{P}_{\boldsymbol{u}})} \right)} \sum_{\boldsymbol{u}\in\Omega} \mathbf{P}_{\boldsymbol{u}} \left\{ \tanh(\kappa \log \mathbf{P}_{\boldsymbol{u}}) - \frac{\sinh(\kappa \log \mathbf{T}_{\boldsymbol{u}})}{\cosh(\kappa \log \mathbf{P}_{\boldsymbol{u}})} \right\}. \tag{38}$$

The χ-exponential family with $\chi(x) = x\beta \log(x)^{1-1/\beta}$ is called the stretched exponential family. The corresponding deformed exponential and logarithm functions can be obtained as $\mathrm{Log}_\chi[x] = \log(x)^{1/\beta}$ and $\mathrm{Exp}_\chi[x] = \exp(x^\beta)$, respectively, and the χ-divergence is defined as

$$D_\beta(\mathbf{T}, \mathbf{P}) = \frac{1}{\sum_{\boldsymbol{u}\in\Omega} \mathbf{T}_{\boldsymbol{u}} \log(\mathbf{T}_{\boldsymbol{u}})^{1-1/\beta}} \sum_{\boldsymbol{u}\in\Omega} \mathbf{T}_{\boldsymbol{u}} \log(\mathbf{T}_{\boldsymbol{u}})^{1-1/\beta} \left\{ \log(\mathbf{T}_{\boldsymbol{u}})^{1/\beta} - \log(\mathbf{P}_{\boldsymbol{u}})^{1/\beta} \right\}.$$

F.4 χ-Fisher information matrix

It can be verified that, when $\chi(x) = x^q$, the χ-Fisher information matrix in Equation (25) coincides with the q-Fisher information matrix (Amari, 2021)

$$\mathbf{G}(\boldsymbol{\theta})_{\boldsymbol{vw}} = \frac{q}{\sum_{\boldsymbol{u}\in\Omega} \mathbf{P}_{\boldsymbol{u}}^q} \sum_{\boldsymbol{u}\in\Omega} \mathbf{P}_{\boldsymbol{u}}^{2q-1} \left(\zeta(\boldsymbol{v}, \boldsymbol{u}) - \eta_{\boldsymbol{v}} \right) \left(\zeta(\boldsymbol{w}, \boldsymbol{u}) - \eta_{\boldsymbol{w}} \right).$$

Furthermore, when $q = 1$, it coincides with the standard Fisher information matrix in the log-linear model on poset, which is derived by Sugiyama et al. (2017).

$$\begin{aligned} \mathbf{G}(\boldsymbol{\theta})_{\boldsymbol{vw}} &= \sum_{\boldsymbol{u}\in\Omega} \mathbf{P}_{\boldsymbol{u}} \left(\zeta(\boldsymbol{v}, \boldsymbol{u}) - \eta_{\boldsymbol{v}} \right) \left(\zeta(\boldsymbol{w}, \boldsymbol{u}) - \eta_{\boldsymbol{w}} \right) \\ &= \sum_{\boldsymbol{u}\in\Omega} \mathbf{P}_{\boldsymbol{u}} \zeta(\boldsymbol{v}, \boldsymbol{u}) \zeta(\boldsymbol{w}, \boldsymbol{u}) - \eta_{\boldsymbol{w}} \eta_{\boldsymbol{v}}, \end{aligned} \tag{39}$$

where we use the normalizing condition $\sum_{\boldsymbol{u}\in\Omega} \mathbf{P}_{\boldsymbol{u}} = 1$. For Kaniadakis statistics, where $\chi(x) = x/\cosh(\kappa \log x)$, the Fisher information matrix is explicitly given as

$$\mathbf{G}(\boldsymbol{\theta})_{\boldsymbol{vw}} = \frac{1}{\sum_{\boldsymbol{u}\in\Omega} \left(\frac{\mathbf{P}_{\boldsymbol{u}}}{\cosh(\kappa \log \mathbf{P}_{\boldsymbol{u}})} \right)} \sum_{\boldsymbol{u}\in\Omega} \mathbf{P}_{\boldsymbol{u}} \frac{1 - \kappa \tanh(\kappa \log \mathbf{P}_{\boldsymbol{u}})}{\{\cosh(\kappa \log \mathbf{P}_{\boldsymbol{u}})\}^2} \left(\zeta(\boldsymbol{v}, \boldsymbol{u}) - \eta_{\boldsymbol{v}} \right) \left(\zeta(\boldsymbol{w}, \boldsymbol{u}) - \eta_{\boldsymbol{w}} \right).$$

The standard Fisher information matrix in Equation (39) is recovered when $\kappa \to 0$.

G REMARKS

The following presents additional remarks that could not be included in the main text due to space limitations.

G.1 Perturbation analysis for further understanding of the deformed models

To elucidate the distinct behaviors of the Tsallis-deformed many-body approximations (q-MBA) and Kaniadakis-deformed many-body approximations (κ-MBA), we investigate how the χ-divergence reacts locally to a perturbation in the tensor. Specifically, we show below that in the κ-MBA, the χ-divergence is more sensitive than the q-MBA to perturbations. For simplicity, the discussion below focuses on the one-body case with $D = 2$, while the extension to higher-order and higher-body cases is straightforward.

We consider $(\chi, 1)$-body tensor $\mathbf{P} \in \mathscr{B}_\chi^{\leq 1}$ and its $O_p(\sigma)$ perturbation ϵ that defines $\mathbf{Q}^\epsilon \in \mathscr{B}_\chi^{\leq 1}$ as

$$\mathbf{P}_{ij} = \frac{1}{Z_\chi^{\mathbf{P}}} \circledast_\chi u_i \circledast_\chi v_j, \quad \mathbf{Q}_{ij}^\epsilon = \frac{1}{Z_\chi^{\mathbf{Q}^\epsilon}} \circledast_\chi u_i \circledast_\chi \left(v_j + \epsilon \delta v_j \right), \tag{40}$$

for arbitrary non-negative vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^M$, and normalizers $Z_\chi^{\mathbf{P}}$ and $Z_\chi^{\mathbf{Q}^\epsilon}$. The perturbation ϵ follows a zero centered Gaussian distribution $\mathcal{N}(0, \sigma^2)$. We evaluate the χ-divergence $L_\chi(\epsilon) \coloneqq D_\chi(\mathbf{Q}^\epsilon, \mathbf{P})$. Assuming $|\epsilon| \ll 1$, we expand $L_\chi(\epsilon)$ around $\epsilon = 0$ as

$$L_\chi(\epsilon) = L_\chi(0) + \epsilon \frac{d}{d\epsilon} L(\epsilon)\Big|_{\epsilon=0} + \epsilon^2 \frac{d^2}{d\epsilon^2} L(\epsilon)\Big|_{\epsilon=0} + O_p(\epsilon^3).$$

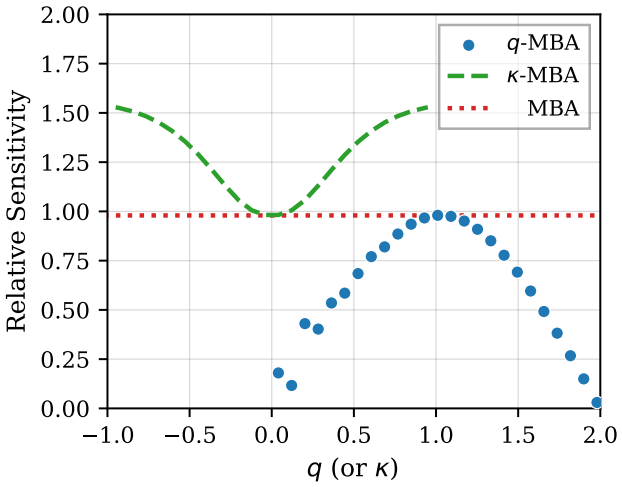


Figure 14: Relative sensitivity, which is the sensitivity divided by that of the ordinary MBA, to remove scaling redundancy. Each element in $\boldsymbol{u}$ and $\boldsymbol{v}$ is independently sampled from a uniform distribution from 0 to 1. The Kaniadakis-deformed many-body approximation (κ-MBA) exhibits higher sensitivity than the standard many-body approximation (MBA), whereas the Tsallis-deformed many-body approximation (q-MBA) exhibits lower sensitivity, corroborating our experimental results in Figures 3, 6 and 12. The parameters κ and q allow control over the sensitivity. Numerical instability around $q = 0$ arises from the definition of the q-algebra.

We assume that the function $L_\chi(\epsilon)$ is three times differentiable. We define the sensitivity to the perturbation as the expectation value over the perturbations, approximated by the average over N finite samples $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$, i.e.,

$$\mathbb{E}_\epsilon[L_\chi(\epsilon) - L_\chi(0)] \simeq \frac{1}{N}\sum_{\ell=1}^{N} L_\chi(\epsilon_\ell) - L_\chi(0) = \frac{1}{N}\sum_{\ell=1}^{N} \epsilon_\ell^2 \frac{d^2}{d\epsilon_\ell^2} L(\epsilon_\ell)\Big|_{\epsilon_\ell=0} + O_p(\rho^3),$$

where $\sum_{\ell=1}^{N} \epsilon_\ell L'(0) \to 0$ $(N \to \infty)$ by Lemma 1, and ρ^3 denotes the third moment of finite perturbations[5], that is $\rho^3 = \mathbb{E}_\epsilon[\epsilon^3]$. We numerically evaluate the sensitivity using $N = 100$ and $M = 100$ in Figure 14, where we vary the parameters q and κ. We use numerical differentiation to approximate the second-order derivative. Each perturbation ϵ_ℓ is independently drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with $\sigma = 10^{-4}$.

The κ-MBA exhibits high sensitivity to perturbations. In particular, larger values of $|\kappa|$ lead to stronger responses, which explains the overfitting to noise observed for large κ in Figure 6 (right) and the smaller slope of the negative log-likelihood in Figure 3. In contrast, the q-MBA shows low sensitivity to perturbations, effectively promoting robust fitting under noise. This accounts for the stronger noise filtering achieved by larger q in Figures 6 (left) and 12.

These observations suggest the following guidelines for model selection in the many-body approximation. For applications that require to reliably capture perturbations or anomalies beyond the standard many-body approximation, the κ-deformed many-body approximation seems to be preferable, with the parameter κ controlling the degree of anomaly detection. Conversely, when we want to suppress contamination in the data, the q-MBA turns out to be advantageous, with the parameter q determining the extent of noise removal.

We also regard perturbations in a tensor as perturbations in its factors in Equation (40), which can be justified by the fact that any tensor $\mathbf{Q}^\epsilon$ is in the manifold $\mathscr{B}_\chi^{\leq m}$ for a sufficiently large m.

G.2 Normalization for deformed many-body approximation

In our proposed Algorithm 1, after updating the natural parameter $\boldsymbol{\theta}$, we need to update the distribution $\mathbf{P}$ to obtain the expectation parameter $\boldsymbol{\eta}$. To obtain the distribution $\mathbf{P}$ using Equation (2), we also need the χ-free

[5]We note for completeness that $L_\chi(0) = 0$ in our experiments

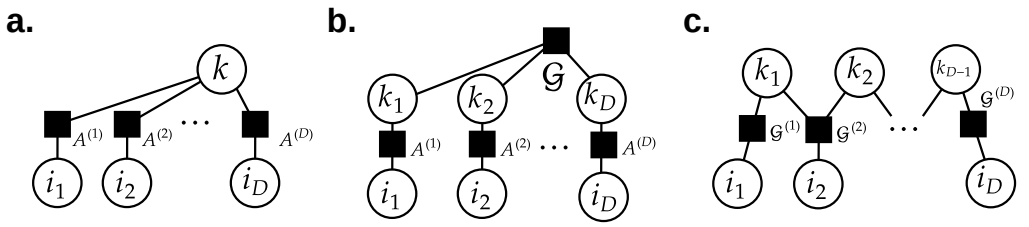


Figure 15: The interactions in the tensor $\mathbf{R}^{[\ell]}$ for $\ell =$ CP (a), $\ell =$ Tucker (b), and $\ell =$ TT (c), where nodes represent the modes and black squares, ■, indicate interactions between the modes.

energy $\psi_\chi(\boldsymbol{\theta})$. However, its closed form is not explicitly available for many χ-functions. Recognizing that the free energy $\psi_\chi(\boldsymbol{\theta})$ serves as a normalization factor, we determine its value by numerically solving the equation

$$\sum_{\boldsymbol{i}\in\Omega} \mathrm{Exp}_\chi\left[-E_{\boldsymbol{i}} - \psi_\chi\right] = 1, \tag{41}$$

for ψ_χ. More specifically, we consider the function

$$g(\psi_\chi) = \sum_{\boldsymbol{i}\in\Omega} \mathrm{Exp}_\chi\left[-E_{\boldsymbol{i}} - \psi_\chi\right] - 1.$$

to obtain the value of ψ_χ. The function g is monotonic since it holds that

$$\frac{\partial}{\partial\psi_\chi} g(\psi_\chi) = -\sum_{\boldsymbol{i}\in\Omega} \chi\left[\mathrm{Exp}_\chi\left[-E_{\boldsymbol{i}} - \psi_\chi\right]\right] = -\sum_{\boldsymbol{i}\in\Omega} \chi\left[\mathbf{P}_{\boldsymbol{i}}\right] < 0,$$

where we use $\lambda(t) = \chi\left[\mathrm{Exp}_\chi\left[t\right]\right]$ and

$$\frac{d}{dt}\mathrm{Exp}_\chi\left[t\right] = \chi\left[\mathrm{Exp}_\chi\left[t\right]\right].$$

Tsallis, stretched, and Kaniadakis-exponential functions satisfy

$$\mathrm{Exp}_\chi\left[-\infty\right] = 0, \quad \mathrm{Exp}_\chi\left[+\infty\right] = \infty.$$

Consequently, it holds that $g(-\infty) = \infty$ and $g(\infty) = -1$. Therefore, the ψ_χ satisfying $g(\psi_\chi) = 0$ exists uniquely and can be found by the bisection method.

We emphasize that the normalization procedure does not become a computational bottleneck, since the bisection method for a univariate monotonic function can be computed extremely efficiently (Hollender et al., 2025).

G.3 Deformed decomposition for various low-rank structure

As seen in Section 3.2, introducing a latent variable allows us to discuss the deformed extension of the CP decomposition (Hitchcock, 1927), which is the most standard low-rank structure in the tensor community. Since tensors have multiple modes, various low-rank structures beyond the CP structure have been defined. By introducing multiple latent variables, we can, in principle, represent arbitrary complex low-rank structures. In the following, we first review the ordinary CP, Tucker (Tucker, 1966), and Tensor Train (TT) (Oseledets, 2011) structures and their connections to the ordinary many-body approximation in Section G.3.1, and then introduce their deformed extensions in Section G.3.2.

G.3.1 CP, Tucker, and TT structures and their connections to many-body approximations

We use the label $\ell \in \{\text{CP, Tucker, TT}\}$ to indicate the low-rank structure, and let $\mathbf{P}^{[\ell]} \in \mathbb{R}^{I_1\times\cdots\times I_D}$ denote a D-th order low-rank tensor whose low-rank structure is specified by the label ℓ. These tensors can be expressed as the marginalization of a $(D+V)$-th order tensor $\mathbf{R} \in \mathbb{R}^{I_1\times\cdots\times I_D\times K_1\times\cdots\times K_V}$ as $\mathbf{P}^{[\ell]}_{\boldsymbol{i}} = \sum_{k_1=1}^{K_1}\cdots\sum_{k_V=1}^{K_V}\mathbf{R}^{[\ell]}_{\boldsymbol{ik}}$

where V is 1 for $\ell = \text{CP}$, D for $\ell = \text{Tucker}$, and $D-1$ for $\ell = \text{TT}$ and the subscript $\boldsymbol{ik}$ denotes the concatenation of $\boldsymbol{i} \in \Omega_I = [I_1] \times \cdots \times [I_D]$ and $\boldsymbol{k} \in \Omega_K = [K_1] \times \cdots \times [K_V]$, i.e., $\boldsymbol{ik} = (i_1, \dots, i_D, k_1, \dots, k_V) \in \Omega_I \times \Omega_K$. Specifically, the higher-order tensor $\mathbf{R}^{[\ell]}$ has the following structure:

$$\mathbf{R}^{[\text{CP}]}_{\boldsymbol{ik}} \coloneqq A^{(1)}_{i_1 k} A^{(2)}_{i_2 k} \dots A^{(D)}_{i_D k}, \tag{42}$$

$$\mathbf{R}^{[\text{Tucker}]}_{\boldsymbol{ik}} \coloneqq G_{k_1 \dots k_D} A^{(1)}_{i_1 k_1} A^{(2)}_{i_2 k_2} \dots A^{(D)}_{i_D k_D}, \tag{43}$$

$$\mathbf{R}^{[\text{TT}]}_{\boldsymbol{ik}} \coloneqq G^{(1)}_{i_1 k_1} G^{(2)}_{k_1 i_2 k_2} G^{(3)}_{k_2 i_3 k_3} \dots G^{(D)}_{k_{D-1} i_D}, \tag{44}$$

where factor matrices $A^{(d)} \in \mathbb{R}^{I_d \times K}$ for $\ell = \text{CP}$, core tensor $G \in \mathbb{R}^{K_1 \times \cdots \times K_D}$ and factor matrices $A^{(d)} \in \mathbb{R}^{I_d \times K_d}$ for $\ell = \text{Tucker}$, and train cores $G^{(d)} \in \mathbb{R}^{K_{d-1} \times I_d \times K_d}$ for $\ell = \text{TT}$. Finding the tensor $\mathbf{P}^{[\ell]}$ that approximates a given tensor $\mathbf{T}$ is referred to as the CP decomposition, Tucker decomposition, or Tensor Train decomposition, according to the choice of ℓ. The degrees of freedom of the hidden variables $(K_1, \dots, K_V)$, referred to as the CP rank, Tucker rank, and Tensor Train rank, respectively, are hyperparameters controlling the number of model parameters.

The tensor $\mathbf{R}^{[\ell]}$ is a low-body tensor in which interactions among its modes are constrained, and these constraints depend on ℓ. For example, the tensor $\mathbf{R}^{[\text{CP}]}$ includes all one-body interactions and two-body interactions only between the variables $i_1, \dots, i_D$ and the variable k. The interactions among tensor modes can be represented by a factor graph, which is called an interaction diagram. Figure 15 illustrates the interactions among tensor modes used in the tensor $\mathbf{R}^{[\ell]}$ for each ℓ. The above connection between tensor low-rank approximation and the many-body approximation was established in (Ghalamkari et al., 2024).

G.3.2 Deformed extension of CP, Tucker, and Tensor Train structures

We introduce the tensor $\mathbf{R}^{[\chi,\ell]}$, in which the products in Equations (42)–(44) are replaced with χ-deformed products $\circledast$, and define the deformed low-rank tensor $\mathbf{P}^{[\chi,\ell]}$ by marginalization of $\mathbf{R}^{[\chi,\ell]}$ over the variables $(k_1, \dots, k_V)$, that is, $\mathbf{P}^{[\chi,\ell]}_{\boldsymbol{i}} = \sum_{k_1=1}^{K_1} \cdots \sum_{k_V=1}^{K_V} \mathbf{R}^{[\chi,\ell]}_{\boldsymbol{ik}}$ where

$$\mathbf{R}^{[\chi,\text{CP}]}_{\boldsymbol{ik}} \coloneqq A^{(1)}_{i_1 k} \circledast A^{(2)}_{i_2 k} \circledast \dots \circledast A^{(D)}_{i_D k}, \tag{45}$$

$$\mathbf{R}^{[\chi,\text{Tucker}]}_{\boldsymbol{ik}} \coloneqq G_{k_1 \dots k_D} \circledast A^{(1)}_{i_1 k_1} \circledast A^{(2)}_{i_2 k_2} \circledast \dots \circledast A^{(D)}_{i_D k_D}, \tag{46}$$

$$\mathbf{R}^{[\chi,\text{TT}]}_{\boldsymbol{ik}} \coloneqq G^{(1)}_{i_1 k_1} \circledast G^{(2)}_{k_1 i_2 k_2} \circledast G^{(3)}_{k_2 i_3 k_3} \circledast \dots \circledast G^{(D)}_{k_{D-1} i_D}, \tag{47}$$

Then, as in the discussion of Section 4.3, for any tensor

$$\mathbf{Q} \in \mathscr{D} = \left\{ \mathbf{Q} \in \mathbb{R}^{I_1 \times \cdots \times I_D \times K_1 \times \cdots \times K_V} \,\middle|\, \sum_{\boldsymbol{k} \in \Omega_K} \mathbf{Q}_{i_1 \dots i_D k_1 \dots k_D} = \mathbf{T}_{i_1 \dots i_D} \right\},$$

the inequality $D^f(\mathbf{T}, \mathbf{P}^{[\chi,\ell]}) \leq D^f(\mathbf{Q}, \mathbf{R}^{[\chi,\ell]})$ holds and it guarantees that the iterative following e-step and m-step decrease the divergence $D^f(\mathbf{T}, \mathbf{P})$ monotonically.

e-step Fixing tensor $\mathbf{R}^{[\chi,\ell]}$, we minimize the upper bound $D^f(\mathbf{Q}, \mathbf{R}^{[\chi,\ell]})$ for $\mathbf{Q} \in \mathscr{D}$. As seen in proposition 10, regardless of the choice of f, this optimization has a closed-form solution given by

$$\mathbf{Q}^* = \frac{\mathbf{T}_{\boldsymbol{i}} \mathbf{R}^{[\chi,\ell]}_{\boldsymbol{ik}}}{\sum_{\boldsymbol{k}} \mathbf{R}^{[\chi,\ell]}_{\boldsymbol{ik}}}. \tag{48}$$

m-step Fixing tensor $\mathbf{Q}$, we minimize $D^f(\mathbf{Q}, \mathbf{R}^{[\chi,\ell]})$ for $\mathbf{R} \in \mathscr{B}^{[\ell]}_\chi$ where $\mathscr{B}^{[\ell]}_\chi$ is the set of low-body tensors whose interactions are given in Figure 15. This is a χ-deformed many-body approximation and an m-projection onto the e-flat manifold $\mathscr{B}^{[\ell]}_\chi$, thus, a convex optimization problem.

In the above discussion, we only considered representative low-rank structures such as the CP, Tucker, and Tensor Train. However, since any low-rank structure can be interpreted as a many-body decomposition including latent variables, the deformed low-rank approximation introduced above can be straightforwardly extended to arbitrary low-rank structures, such as the tensor tree (Murg et al., 2015), tensor wheel (Wu et al., 2022), tensor ring (Zheng et al., 2021), and fully connected structures (Zhao et al., 2016).

H LIMITATIONS, FUTURE WORK AND OPEN PROBLEMS

Model selection In our proposed χ-deformed many-body approximation, the χ-function can be any positive function. This flexibility raises a natural question: what kind of χ-function should we use for tensor factorization? Although we have demonstrated the advantages of the Tsallis and Kaniadakis deformed many-body approximation in this work, investigating the benefits of other χ-functions or designing tailored χ-functions for specific applications remains an open research direction.

Limitation of the *em*-algorithm The proposed *em*-based deformed low-rank approximation is currently limited to cases where the χ-divergence optimization coincides with that of the f-divergence, and extending it beyond this setting is an important direction for future research. The proposed deformed low-rank approximation exhibits the same issues as the conventional EM algorithm, such as non-convex optimization and convergence to stationary points.

Normalization constraint Our framework assumes normalization of the tensor, as it is based on information geometry. Due to this constraint, each iteration requires a normalization procedure and must access all elements of the tensor, which contrasts with existing tensor approaches scaling by samples, i.e., exploiting sparsity (Ghalamkari et al., 2024; Chege et al., 2022; Yeredor and Haardt, 2019). We expect that developing alternative parameterizations or techniques to overcome this limitation, such as approximating the normalization using sampling procedures, can significantly improve the scalability of the deformed low-rank approximation.

Semantic meaning of the decomposition Compared with the conventional low-rank factorization based on the sum-of-products form, the deformed low-rank approximation provides a less intuitive factorization. For example, although the Tsallis-deformed product acts as an intermediate operation between sum and product, controlled by the parameter q, the semantic meaning of the obtained low-rank structure, which is given as the sum of this operation, is not necessarily clear.

Missing closed-form updates The property of the logarithm function, $\log ab = \log a + \log b$, in the KL-divergence, often enables us to find the closed-form updates for the M-step of the traditional EM-algorithm since the corresponding KL-divergence-based many-body (or rank-1) approximation has closed-form solutions (Murphy, 2012; Ghalamkari et al., 2024; Ghalamkari and Sugiyama, 2022). In contrast, the optimization for χ-divergence requires a gradient-based update in each m-step since the closed-form solution of the corresponding deformed many-body approximation is not trivial. As a result, we need to perform iterative updates in each m-step in the proposed deformed low-rank decomposition, as seen in Algorithm 2. A more efficient update is the focus of future work.