Journal of Machine Learning Research X (20XX) Submitted 00/00; Published 00/00

Learning from Binary Multiway Data: Probabilistic Tensor Decomposition and its Statistical Optimality

Miaoyan Wang MIAOYAN.WANG@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706, USA

Lexin Li LEXINLI@BERKELEY.EDU
Department of Biostatistics and Epidemiology
University of California-Berkeley
Berkeley, CA 94720, USA

Editor: Animashree Anandkumar

Abstract

We consider the problem of decomposition of multiway tensor with binary entries. Such data problems arise frequently in applications such as neuroimaging, recommendation system, topic modeling, and sensor network localization. We propose a multilinear Bernoulli model, develop a rank-constrained likelihood-based estimation, and obtain the theoretical accuracy guarantees. In contrast to continuous-valued problems, the binary tensor problem exhibits an interesting phase change phenomenon according to the signal-to-noise ratio. We establish the error bound of the tensor estimation, and show that the obtained rate is minimax optimal in high dimensions under the considered model. We also develop an alternating updating algorithm with convergence guarantees. The efficacy of our approach is demonstrated through both simulations and analyses of multiple real-world datasets on the tasks of tensor completion and clustering.

Keywords: Binary tensor, CP tensor decomposition, Constrained maximum likelihood estimation, Diverging dimensionality, Generalized linear model.

1. Introduction

1.1 Motivation and proposal

In recent years, data in the form of multidimensional arrays, a.k.a. tensors, are frequently arising and gaining increasing attention in numerous fields, such as genomics (Hore et al., 2016; Wang et al., 2017c), neuroscience (Zhou et al., 2013), recommender systems (Sun et al., 2017; Bi et al., 2018), social networks (Nickel et al., 2011), computer vision (Tang et al., 2013), among many others. An important reason is the effective representation of multiway data using a tensor structure. One example is the recommender system (Bi et al., 2018), which can be naturally described as a three-way tensor of user $\times$ item $\times$ context and each entry indicates the user-item interaction. Another example is the DBLP database (Zhe et al., 2016), which is organized into a three-way tensor of author $\times$ word $\times$ venue and each entry indicates the co-occurrence of the triplets.

Whereas many real-world multiway datasets have continuous-valued entries, there have recently emerged more instances of binary tensors, in which all tensor entries are binary indicators 0/1. Examples include click/no-click action in recommender systems (Sun et al., 2017), multi-relational social networks (Nickel et al., 2011), and brain structural connectivity networks (Wang et al., 2017a). These binary tensors are often noisy and high-dimensional. It is thus crucial to develop effective tools that reduce the dimensionality, take into account the tensor formation, and learn the underlying structures of these massive discrete observations. A number of successful tensor decomposition methods have been proposed (Kolda and Bader, 2009; Anandkumar et al., 2014; Wang and Song, 2017), revitalizing the classical methods such as CANDECOMP/PARAFAC (CP) decomposition (Hitchcock, 1927) and Tucker decomposition (Tucker, 1966), as well as developing new ones such as tensor train decomposition (Oseledets, 2011). These methods treat tensor entries as continuous-valued, and are therefore not suitable to handle binary tensors. There is a relative paucity of binary tensor decomposition methods.

In this article, we develop a general method and the associated theory for binary tensor decomposition. For a binary tensor $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \{0,1\}^{d_1\times\cdots\times d_K}$, whose entries are either 1 or 0 that encodes the presence or absence of the event indexed by the K-tuplet $(i_1,\ldots,i_K)$, we propose to consider the following low-rank Bernoulli model:

$$\mathcal{Y}|\Theta \sim \text{Bernoulli}\{f(\Theta)\}, \quad \text{where} \quad \text{rank}(\Theta) = R. \tag{1}$$

That is, we assume the entries of $\mathcal{Y}$ are realizations of independent Bernoulli random variables with success probability $f(\theta_{i_1,\ldots,i_K})$, where f is a suitable function that maps $\mathbb{R}$ to $[0,1]$. The parameter tensor, $\Theta = [\![\theta_{i_1,\ldots,i_K}]\!]$ is of the same dimension as $\mathcal{Y}$ but its entries are continuous-valued, and we assume it admits a low-rank CP structure. Our goal is to estimate Θ from the large binary tensor $\mathcal{Y}$ of arbitrary order K. In particular, we are interested in the setting where the dimensions $(d_1,\ldots,d_K)$ diverge. As the tensor dimensions grow, it is crucial to understand both statistical and computational properties, which is to be our primary focus. Specifically, we aim to study some fundamental issues of binary tensor decomposition, including the signal-to-noise ratio under which a reliable estimation of Θ is possible, the intrinsic hardness, in terms of minimax rate, of the binary problem compared to its continuous-valued counterpart, and the estimation properties that are specific to a particular algorithm, and those general to all algorithms.

1.2 Related work

Our work is closely related to but also clearly distinctive from several lines of existing research. We next survey the main related approaches.

Continuous-valued tensor decomposition. In principle, one can apply the existing decomposition methods that were designed for continuous-valued tensor (Kolda and Bader, 2009; Wang and Song, 2017; Zhang and Xia, 2018) to binary tensor, by pretending the 0/1 entries were continuous. However, such an approach is bound to yield an inferior performance: flipping the entry coding $0 \leftrightarrow 1$ would totally change the decomposition result, and the predicted values for the unobserved entries could fall outside the valid range $[0,1]$. Our method, however, is invariant to flipping, as reversing the entry coding of $\mathcal{Y}$ only changes the sign, but not the low-rank structure, of the parameter Θ. Moreover, as we show in Section 3.3, binary tensor decomposition exhibits a "dithering" effect (Davenport et al., 2014)

that necessitates the presence of stochastic noise in order to estimate Θ. This is clearly contrary to the behavior of continuous-valued tensor decomposition.

Binary matrix decomposition. When $K = 2$, the problem reduces to binary or logit principal component analysis (PCA), and a similar model as (1) has been proposed (Collins et al., 2002; De Leeuw, 2006; Lee et al., 2010). While tensors are conceptual generalization of matrices, matrix decomposition and tensor decomposition are fundamentally different (Kolda and Bader, 2009). Under the matrix case, the rank R is required to be no greater than $\min(d_1, d_2)$, and the factor matrices involved are constrained to be orthogonal for the identification purpose. Such constraints, however, are unnecessary for tensors, since the uniqueness of tensor CP decomposition holds under much milder conditions (Bhaskara et al., 2014). Besides, tensors do not always admit orthogonal decomposition, and the tensor rank R can exceed the dimension. These differences make the earlier algorithms that built upon matrix decomposition unsuitable to tensors. Moreover, as we show in Section 3.1, if we were to apply the matrix version of binary decomposition to a tensor by unfolding the tensor into a matrix, the result is suboptimal with a slower convergence rate.

Binary tensor decomposition. More recently, Mažgut et al. (2014); Rai et al. (2015); Hong et al. (2018) studied higher-order binary tensor decomposition, and we target the same problem. However, our study differs in terms of the scope of the results. In general, there are two types of properties that an estimator possesses. The first type is the algorithm-dependent property that quantifies the impact of a specific algorithm, such as the choice of loss function, initialization, and iterations, on the final estimator. The second type is the statistical property that characterizes the population behavior and is independent of any specific algorithm. Earlier solutions of Mažgut et al. (2014); Rai et al. (2015); Hong et al. (2018) focused only on the algorithm effectiveness, but did not address the population optimality. By contrast, we study both types of properties in Sections 3 and 4. This allows us to better understand the gap between a specific algorithm and the population optimality, which may in turn offer a useful guide to the algorithm design.

1-bit completion. Our work is also connected to 1-bit matrix completion (Cai and Zhou, 2013; Davenport et al., 2014) and its recent extension to 1-bit tensor completion (Ghadermarzy et al., 2018). The completion problem aims to recover a matrix or tensor from incomplete observations of its entries. The observed entries are highly quantized, sometimes even to a single bit. Thus the problem turns to recover a signal matrix or tensor based on the noise-corrupted signs of a subset of its entries. We first show in Section 2.2 that our Bernoulli tensor model has an equivalent interpretation as the threshold model commonly used in 1-bit quantization. The two methods are compared in Section 3.1. Assuming the signal rank is known or otherwise can be estimated, we are able to achieve a faster convergence rate than that in 1-bit tensor completion (Ghadermarzy et al., 2018). The optimality of our estimator is safeguarded by a matching minimax lower bound.

Boolean tensor decomposition. Boolean tensor decomposition (Miettinen, 2011; Erdos and Miettinen, 2013a; Rukat et al., 2018; Hu et al., 2018) is a data-driven algorithm that decomposes a binary tensor into binary factors. The idea is to use logic operations to replace addition and multiplication in the factorization. These methods also dealt with binary tensor, same as we do, but they took a model-free approach to approximate the data instance. One important difference is that we focus on parameter estimation in a population model. The population interpretation offers useful insight on the effectiveness of dimension

reduction. In addition, having a population model allows us to tease apart the algorithmic error versus the statistical error. In this respect, our proposal is very different from boolean tensor decomposition. We also numerically compare the two approaches in Section 5.

Bayesian binary tensor decomposition. There have been a number of Bayesian binary tensor decomposition algorithms (Nickel et al., 2011; Rai et al., 2014, 2015). Most of those focused on and were tailored to the specific context of multi-relational learning. Although we take multi-relational learning as one of our applications, we aim to address a general binary tensor decomposition problem, and to study some intrinsic statistical properties of the problem, such as the SNR phase diagram and minimax rate. Besides, we provide a frequentist-type solution which is computationally more tractable than a Bayesian one.

1.3 Contributions

We summarize our main contributions that set our work apart from the existing literature.

First, we systematically quantify the hardness of the binary tensor decomposition problem. We show that the Bernoulli tensor model (1) is equivalent to entrywise quantization of a latent noisy continuous-valued tensor. We then characterize the impact of latent signal-to-noise ratio (SNR) on the tensor recovery accuracy, and identify three different phases for tensor recovery according to SNR; see Table 1 in Section 3.3. When SNR is bounded by a constant, the loss in binary tensor decomposition is comparable to the case of continuous-valued tensor, suggesting very little information has been lost by quantization. On the other hand, when SNR is sufficiently large, stochastic noise turns out to be helpful, and is in fact essential, for estimating the signal tensor. The later effect is related to "dithering" (Davenport et al., 2014) and "perfect separation" (Albert and Anderson, 1984) phenomenon, and is clearly contrary to the behavior of continuous-valued tensor decomposition.

Second, we propose a new method for binary tensor decomposition and establish its statistical properties, including the upper bound and the minimax lower bound on the tensor recovery accuracy. These properties characterize the intrinsic population optimality of the estimator and are independent of any specific algorithm. Note that, in our problem, the tensor dimensions $(d_1, \ldots, d_K)$ diverge, and so does the number of unknown parameters. As such, the classical maximum likelihood estimation (MLE) theory does not directly apply. We leverage the recent development in random tensor theory and high-dimensional statistics, and establish the error bounds of the tensor estimation. The matching information-theoretical lower bounds are correspondingly provided. In particular, when the tensor dimensions are the same in all modes, $d_1 = \ldots = d_K = d$, we obtain a convergence rate $\asymp d^{-(K-1)/2}$ for estimating Θ. This rate outperforms the rate of "best matrixization", which is $\asymp d^{-\frac{\lfloor K/2 \rfloor}{2}}$, and $\lfloor K/2 \rfloor$ is the integer part of $K/2$, as well as the rate of 1-bit tensor completion, which is $\asymp d^{-(K-1)/4}$ (Ghadermarzy et al., 2018). To our knowledge, these statistical guarantees are among the first for binary tensor decomposition.

Lastly, we supplement the above general statistical properties by proposing a block relaxation algorithm, and establish the corresponding algorithmic convergence properties. Our algorithm-dependent error bound reveals an interesting interplay between the computational efficiency and the statistical convergence. We also illustrate the efficacy of our algorithm through both simulations and real data applications.

1.4 Notation and organization

We adopt the following notation throughout the article. We use $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \mathbb{F}^{d_1 \times \cdots \times d_K}$ to denote an order-K $(d_1, \ldots, d_K)$-dimensional tensor over a filed $\mathbb{F}$. We focus on real or binary tensors, i.e., $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \{0, 1\}$. The Frobenius norm of $\mathcal{Y}$ is defined as $\|\mathcal{Y}\|_F = (\sum_{i_1,\ldots,i_K} y^2_{i_1,\ldots,i_K})^{1/2}$, and the infinity norm of $\mathcal{Y}$ is defined as $\|\mathcal{Y}\|_\infty = \max_{i_1,\ldots,i_K} |y_{i_1,\ldots,i_K}|$. We use uppercase letters, such as Θ, $\mathcal{Y}$, $\boldsymbol{A}$, to denote tensors and matrices, and use lowercase letters, such as θ, $\boldsymbol{a}$, to denote scales and vectors. We use $\boldsymbol{a} \otimes \boldsymbol{b}$ to denote the kronecker product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, and $\boldsymbol{A} \odot \boldsymbol{B}$ to denote the Khatri-Rao product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. We use $\boldsymbol{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$ to denote the $(d-1)$-dimensional unit sphere, and the shorthand n-set $\{1, ..., n\}$ to denote $n \in \mathbb{N}_+$.

The rest of the article is organized as follows. We present the low-rank Bernoulli tensor model, its connection with 1-bit observation model, and the rank-constrained MLE framework in Section 2. We establish the statistical properties of the upper and lower bounds and the phase-transition in Section 3. We next develop an alternating updating algorithm and establish its convergence guarantees in Section 4. We present the simulations in Section 5, and the real-world data analyses in Section 6. We conclude the paper with a discussion in Section 7. All technical proofs are deferred to the Appendix.

2. Model

2.1 Low-rank Bernoulli model

For the binary tensor $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \{0, 1\}^{d_1 \times \cdots \times d_K}$, we assume its entries are realizations of independent Bernoulli random variables, such that, for all $(i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K]$,

$$\mathbb{P}(y_{i_1,\ldots,i_K} = 1) = f(\theta_{i_1,\ldots,i_K}). \tag{2}$$

In this model, $f : \mathbb{R} \to [0, 1]$ is a strictly increasing function. We further assume that $f(\theta)$ is twice-differentiable in $\theta \in \mathbb{R}/\{0\}$; $f(\theta)$ is strictly increasing and strictly log-concave; and $f'(\theta)$ is unimodal and symmetric with respect to $\theta = 0$. All these assumptions are fairly mild. In the language of generalized linear model (GLM), f is often referred to as the "inverse link function". When no confusion arises, we also call f the "link function". The parameter tensor $\Theta = [\![\theta_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is continuous-valued, unknown, and is the main object of interest in our tensor estimation inquiry. We also assume that the entries of $\mathcal{Y}$ are mutually independent conditional on Θ, which is again commonly adopted in the literature (Collins et al., 2002; De Leeuw, 2006; Lee et al., 2010). Note that this assumption does not rule out the marginal correlations among the entries of $\mathcal{Y}$.

Furthermore, we assume the parameter tensor Θ admits a rank-R CP decomposition,

$$\Theta = \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)}, \tag{3}$$

where $\lambda_r \in \mathbb{R}_+$, with $\lambda_1 \geq \ldots \geq \lambda_K$ without loss of generality, $\boldsymbol{a}_r^{(k)} \in \mathbf{S}^{d_k - 1}$, for $r \in [R], k \in [K]$, and Θ cannot be written as a sum of fewer than R outer products. The CP structure in (3) is frequently used in tensor data analysis, and is shown to provide a reasonable tradeoff between model complexity and model flexibility (Chen et al., 2016). Moreover,

this structure seems plausible in numerous scientific applications. For instance, in sensor network locations, DNA haplotype assembly, and brain imaging analysis, it is often found that the rank of Θ is small, is independent of the tensor dimension, and offers a reasonable approximation to the truth (Zhou et al., 2013; Bhaskar and Javanmard, 2015). It helps dramatically reduce the number of parameters in Θ, from the order of $\prod_k d_k$ to the order of $\sum_k d_k$. More precisely, the effective number of parameters in (3) is $p_e = R(d_1 + d_2) - R^2$ for $K = 2$ matrices after adjusting for the nonsingular transformation indeterminacy, and $p_e = R(\sum_k d_k - K + 1)$ for $K \geq 3$ higher-order tensors after adjusting for the scaling indeterminacy (Zhou et al., 2013).

Combining (2) and (3) leads to our low-rank Bernoulli model. We seek to estimate the rank-R tensor Θ given the observed binary tensor $\mathcal{Y}$. The model can be viewed as a generalization of the classical CP decomposition for continuous-valued tensors to binary tensors, in a way that is analogous to the generalization from a linear model to a GLM. When imposing this structure to a continuous-valued tensor $\mathcal{Y}$ directly, it amounts to seek the best rank-R approximation to $\mathcal{Y}$, in the least square sense. Correspondingly, the least-square criterion can be interpreted as a maximum likelihood procedure for finding a low-rank tensor Θ from a noisy observation $\mathcal{Y} = \Theta + \mathcal{E}$, where $\mathcal{E} \in \mathbb{R}^{d1} \times \cdots \times \mathbb{R}^{d_k}$ collects i.i.d. Gaussian noises. In the next section, we see a close connection between a real-valued tensor problem and a binary tensor problem.

2.2 Latent variable model interpretation

We next show that our binary tensor model (2) has an equivalent interpretation as the threshold model commonly used in 1-bit quantization (Davenport et al., 2014; Bhaskar and Javanmard, 2015; Cai and Zhou, 2013; Ghadermarzy et al., 2018). The later viewpoint sheds light on the nature of the binary (1-bit) measurements from the information perspective.

Consider an order-K tensor $\Theta = [\![\theta_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ with a rank-R CP structure. Suppose that we cannot not directly observe Θ. Instead, we only observe the quantized version $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \{0,1\}^{d_1 \times \cdots \times d_K}$ following the scheme

$$y_{i_1,\ldots,i_K} = \begin{cases} 1 & \text{if } \theta_{i_1,\ldots,i_K} + \varepsilon_{i_1,\ldots,i_K} \geq 0, \\ 0 & \text{if } \theta_{i_1,\ldots,i_K} + \varepsilon_{i_1,\ldots,i_K} < 0, \end{cases} \tag{4}$$

where $\mathcal{E} = [\![\varepsilon_{i_1,\ldots,i_K}]\!]$ is a noise tensor. That is, the binary tensor is generated from $\mathcal{Y} = \text{sign}(\Theta + \mathcal{E})$, and the associated latent tensor is $\Theta + \mathcal{E}$. Here the sign function $\text{sign}(x) \overset{\text{def}}{=} \mathbb{1}_{\{x \geq 0\}}$ is applied to tensors in an element-wise manner. From the viewpoint of (4), the tensor Θ is not merely an argument in the Bernoulli tensor model (2); it can be interpreted as an underlying, continuous-valued quantity whose noisy discretization gives $\mathcal{Y}$.

The latent model (4) in fact is equivalent to our Bernoulli tensor model (2), if the link f behaves like a cumulative distribution function. Specifically, for any choice of f in (2), if we define $\mathcal{E}$ as having i.i.d. entries drawn from a distribution whose cumulative distribution function is defined by $\mathbb{P}(\varepsilon < \theta) = 1 - f(-\theta)$, then (2) reduces to (4). Conversely, if we set the link function $f(\theta) = \mathbb{P}(\varepsilon \geq -\theta)$, then model (4) reduces to that in (2). There is a one-to-one correspondence between the error distribution in the latent model and the link function in the Bernoulli model. We next consider three choices of f, or equivalently, the distribution of $\mathcal{E}$.

Example 1. (Logistic link/Logistic noise). The logistic model is represented by (2) with $f(\theta) = \left(1 + e^{-\theta/\sigma}\right)^{-1}$ and the scale parameter $\sigma > 0$. Equivalently, the noise $\varepsilon_{i_1,\ldots,i_K}$ in (4) follows i.i.d. logistic distribution with the scale parameter σ.

Example 2. (Probit link/Gaussian noise). The probit model is represented by (2) with $f(\theta) = \Phi(\theta/\sigma)$, where Φ is the cumulative distribution function of a standard Gaussian. Equivalently, the noise $\varepsilon_{i_1,\ldots,i_K}$ in (4) follows i.i.d. $N(0, \sigma^2)$.

Example 3. (Laplacian link/Laplacian noise). The Laplacian model is represented by (2) with

$$f(\theta) = \begin{cases} \frac{1}{2}\exp\left(\frac{\theta}{\sigma}\right), & \text{if } \theta < 0, \\ 1 - \frac{1}{2}\exp(-\frac{\theta}{\sigma}), & \text{if } \theta \geq 0, \end{cases}$$

and the scale parameter $\sigma > 0$. Equivalently, the noise $\varepsilon_{i_1,\ldots,i_K}$ in (4) follows i.i.d. Laplace distribution with the scale parameter σ.

The above link functions are common for the Bernoulli model, and the choice is informed by several considerations. The probit is the canonical link based on the Bernoulli likelihood, and has a direct connection with the log-odds of success. The probit is connected to thresholded latent Gaussian tensors. The Laplace has a heavier tail than the normal distribution, so is often more suitable for modeling long-tail data.

2.3 Rank-constrained likelihood-based estimation

We next propose to estimate the unknown parameter tensor Θ in model (2) using a constrained likelihood approach. The log-likelihood function for (2) is

$$\begin{aligned} \mathcal{L}_{\mathcal{Y}}(\Theta) &= \sum_{i_1,\ldots,i_K} \left[\mathbb{1}_{\{y_{i_1,\ldots,i_K}=1\}} \log f(\theta_{i_1,\ldots,i_K}) + \mathbb{1}_{\{y_{i_1,\ldots,i_K}=0\}} \log\left\{1 - f(\theta_{i_1,\ldots,i_K})\right\} \right] \\ &= \sum_{i_1,\ldots,i_K} \log f(q_{i_1,\ldots,i_K}\theta_{i_1,\ldots,i_K}), \end{aligned}$$

where $q_{i_1,\ldots,i_K} = 2y_{i_1,\ldots,i_K} - 1$ takes values -1 or 1, and the second equality is due to the symmetry of the link function f. To incorporate the CP structure (3), we consider a constrained optimization:

$$\max_{\Theta \in \mathcal{D}} \mathcal{L}_{\mathcal{Y}}(\Theta), \quad \text{where } \mathcal{D} \subset \mathcal{S} = \{\Theta : \operatorname{rank}(\Theta) = R, \text{ and } \|\Theta\|_\infty \leq \alpha\}, \tag{5}$$

for a given rank $R \in \mathbb{N}_+$ and a bound $\alpha \in \mathbb{R}_+$. Here the search space $\mathcal{D}$ is assumed to be a compact set containing the true parameter Θ_{true}. The candidate tensor of our interest satisfies two constraints. The first is that Θ admits the CP structure (3) with rank R. As discussed in Section 2.1, the low-rank structure (3) is an effective dimension reduction tool in tensor data analysis. The second constraint is that all the entries of Θ are bounded in absolute value by a constant $\alpha \in \mathbb{R}_+$. We refer to α as the "signal" bound of Θ. This infinity-norm condition is a technical assumption to aid the recovery of Θ in the noiseless case. Similar techniques have been employed for the matrix case (Davenport et al., 2014; Bhaskar and Javanmard, 2015; Cai and Zhou, 2013).

In the next section, we first investigate the statistical properties of the global constrained maximum likelihood estimator, $\hat{\Theta}_{\text{MLE}} = \arg\max_{\Theta\in\mathcal{D}} \mathcal{L}_{\mathcal{Y}}(\Theta)$, in terms of error bounds for the estimation accuracy. These bounds characterize the population behavior of the global estimator, and weave three quantities: tensor dimension, rank, and signal-to-noise ratio. We then compare these properties to the information-theoretical bound and reveal an intrinsic phase-transition phenomenon. In Section 4, we develop a specific algorithm for the optimization problem in (5), and derive the convergence properties of the actual estimator resulting from this algorithm.

3. Statistical Properties

3.1 Performance upper bound

We define two quantities L_α and γ_α to control the "steepness" and "convexity" of the link function f. Let

$$L_\alpha = \sup_{|\theta|\le\alpha} \left\{ \frac{\dot{f}(\theta)}{f(\theta)\left(1-f(\theta)\right)} \right\}, \quad \text{and} \quad \gamma_\alpha = \inf_{|\theta|\le\alpha} \left\{ \frac{\dot{f}^2(\theta)}{f^2(\theta)} - \frac{\ddot{f}(\theta)}{f(\theta)} \right\},$$

where $\dot{f}(\theta) = df(\theta)/d\theta$, and α is the bound on the entry-wise infinity-norm of Θ. When α is a fixed constant and f is a fixed function, all these quantities are bounded by some fixed constants independent of the tensor dimension. In particular, for the logistic, probit and Laplacian models, we have

$$\begin{aligned}
&\text{Logistic model:} && L_\alpha = \frac{1}{\sigma}, \quad \gamma_\alpha = \frac{e^{\alpha/\sigma}}{(1+e^{\alpha/\sigma})^2\sigma^2}, \\
&\text{Probit model:} && L_\alpha \le \frac{2}{\sigma}\left(\frac{\alpha}{\sigma}+1\right), \quad \gamma_\alpha \ge \frac{1}{\sqrt{2\pi}\sigma^2}\left(\frac{\alpha}{\sigma}+\frac{1}{6}\right)e^{-x^2/\sigma^2}, \\
&\text{Laplacian model:} && L_\alpha \le \frac{2}{\sigma}, \quad \gamma_\alpha \ge \frac{e^{-\alpha/\sigma}}{2\sigma^2}.
\end{aligned}$$

We assess the estimation accuracy using the deviation in the Frobenius norm. For the true coefficient tensor $\Theta_{\text{true}} \in \mathbb{R}^{d_1\times\cdots\times d_K}$ and its estimator $\hat{\Theta}$, define

$$\text{Loss}(\hat{\Theta}, \Theta_{\text{true}}) = \frac{1}{\sqrt{\prod_k d_k}} \|\hat{\Theta} - \Theta_{\text{true}}\|_F.$$

The next theorem establishes the upper bound for $\hat{\Theta}_{\text{MLE}}$ under model (2).

Theorem 1 (Statistical convergence). *Suppose $\mathcal{Y} \in \{0,1\}^{d_1\times\cdots\times d_K}$ is an order-K binary tensor following model (2), with the link function f, and the true coefficient tensor $\Theta_{true} \in \mathcal{D}$. Then there exist an absolute constant $C_1 > 0$, and a constant $C_2 > 0$ that depends only on K, such that, with probability at least $1 - \exp\left(-C_1 \log K \sum_k d_k\right)$,*

$$Loss(\hat{\Theta}_{MLE}, \Theta_{true}) \le \min\left(2\alpha, \ \frac{C_2 L_\alpha}{\gamma_\alpha}\sqrt{\frac{R^{K-1}\sum_k d_k}{\prod_k d_k}}\right). \tag{6}$$

Its proof is given in the Appendix. Note that f is strictly log-concave if and only if $\ddot{f}(\theta)f(\theta) < \dot{f}(\theta)^2$ (Boyd and Vandenberghe, 2004). Henceforth, $\gamma_\alpha > 0$ and $L_\alpha > 0$, which ensures the validity of the bound in (6). A matching lower bound is given in Section 3.2.

To gain further insight into this upper bound, we consider a special setting where the dimensions are the same in all modes, i.e., $d_1 = \ldots = d_K = d$. In such a case, our bound (6) reduces to

$$\text{Loss}(\hat{\Theta}_{\text{MLE}}, \Theta_{\text{true}}) \leq O\left(\frac{1}{d^{(K-1)/2}}\right),$$

for a fixed rank R and a signal bound α, as $d \to \infty$. The MLE thus achieves consistency with inverse-polynomial convergence rate. Comparing to the error bound for 1-bit tensor recovery (Ghadermarzy et al., 2018),

$$\text{Loss}(\hat{\Theta}, \Theta_{\text{true}}) \leq O\left(\frac{1}{d^{(K-1)/4}}\right),$$

we see that our bound has a faster convergence rate. This rate improvement comes from the fact that we impose an exact low-rank structure on Θ, whereas Ghadermarzy et al. (2018) employed the max norm as a surrogate rank measure.

Our bound also generalizes the previous results on low-rank binary matrix completion. The convergence rate for rank-constrained matrix completion was $O(1/\sqrt{d})$ (Bhaskar and Javanmard, 2015), which fits into our special case when $K = 2$. Intuitively, for tensor data, we can view each tensor entry as a data point, and the total number of entries would correspond to the sample size. Compared to the matrix case, a higher-order tensor has a larger number of data points and thus exhibits a faster convergence rate as $d \to \infty$.

As an immediate corollary of Theorem 1, we obtain the explicit form of the upper bound (6) when the link f is a logistic, probit, or Laplacian function.

Corollary 1. *Assume the same setup as in Theorem 1. Then there exist an absolute constant $C' > 0$ such that with probability at least $1 - \exp(-C' \log K \sum_k d_k)$,*

$$\textit{Loss}(\hat{\Theta}_{\textit{MLE}}, \Theta_{\textit{true}}) \leq \min\left\{2\alpha,\ C(\sigma,\alpha)\sqrt{\frac{R^{K-1}K\sum_k d_k}{\prod_k d_k}}\right\}, \tag{7}$$

where $C(\alpha, \sigma)$ is a scaler factor, and

$$C(\alpha,\sigma) = \begin{cases} C_1\sigma\left(2 + e^{\frac{\alpha}{\sigma}} + e^{-\frac{\alpha}{\sigma}}\right) & \textit{for the logistic link,} \\ C_2\sigma\left(\dfrac{\alpha+\sigma}{6\alpha+\sigma}\right)e^{\frac{\alpha^2}{\sigma^2}} & \textit{for the probit link,} \\ C_3\alpha e^{\frac{\alpha}{\sigma}} & \textit{for the Laplacian link,} \end{cases}$$

and $C_1, C_2, C_3 > 0$ are constants that depend only on K.

The dependency of the above error bounds on the signal bound α and the noise level σ will be investigated in Section 3.3.

3.2 Information-theoretical lower bound

We next establish two lower bounds. The first lower bound is for any statistical estimator $\hat{\Theta}$, including but not necessarily the constrained MLE $\hat{\Theta}_{\text{MLE}}$, under the binary tensor model (2). The result is based on the information theory and is algorithm-independent. It reveals the fundamental hardness of the problem. We show that this lower bound nearly matches the upper bound on the estimation accuracy of $\hat{\Theta}_{\text{MLE}}$, and thus establishes the rate optimality of $\hat{\Theta}_{\text{MLE}}$. With a little abuse of notation, we use $\mathcal{D}(R, \alpha)$ to denote the set of tensors with the rank bounded by R and the infinity norm bounded by α. The next theorem establishes this first lower bound for any estimator $\hat{\Theta}$ in $\mathcal{D}(R, \alpha)$ under model (2).

Theorem 2 (Minimax lower bound for binary tensors). *Suppose $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$ is an order-K binary tensor following model (2), with a probit link function f, and the true coefficient tensor $\Theta_{true} \in \mathcal{D}(R, \alpha)$. Suppose that $R \leq \min_k d_k$ and the dimension $\max_k d_k \geq 8$. Let $\hat{\Theta}$ denote an estimator of Θ_{true}. Then there exist absolute constants $\beta_0 \in (0,1)$ and $c_0 > 0$, such that*

$$\inf_{\hat{\Theta}} \sup_{\Theta_{true} \in \mathcal{D}(R,\alpha)} \mathbb{P}\left\{ Loss(\hat{\Theta}, \Theta_{true}) \geq c_0 \min\left(\alpha,\ \sigma\sqrt{\frac{R d_{\max}}{\prod_k d_k}} \right) \right\} \geq \beta_0. \tag{8}$$

The proof of this theorem is given in the Appendix. Here we only present the result for the probit model, while similar results can be obtained for the logistic and Laplacian model as well. We comment that, in this theorem, we assume that $R \leq d_{\min}$. This condition is automatically satisfied in the matrix case, since the rank of a matrix is always bounded by its row and column dimension. For the tensor case, this assertion may not always hold. However, in the majority applications, the tensor dimension is large, whereas the rank is relatively small. We view this as a mild technical condition. In Section 5, we will assess the empirical performance when the rank exceeds dimension. Also note that in Theorem 1, we do not place any constraint on the rank R.

We next compare the lower bound (8) to the upper bound (7), as the tensor dimension $d_k \to \infty$ while the signal bound α and the noise level σ are fixed. Since $d_{\max} \leq \sum_k d_k \leq K d_{\max}$, both the bounds are of the form $C\sqrt{d_{\max}}\,(\prod_k d_k)^{-1/2}$, where C is a factor that does not depend on the tensor dimension. Henceforth, the estimator $\hat{\Theta}_{\text{MLE}}$ is rate-optimal. For the scenario when the tensor dimensions are the same in all modes, $d_1 = \ldots = d_K = d$, the convergence rate for estimating Θ using our tensor method is $O(d^{-(K-1)/2})$, while the "best" unfolding solution that unfolds a tensor into a near-square matrix (Mu et al., 2014) gives the convergence rate $O(d^{-\frac{\lfloor K/2 \rfloor}{2}})$, with $\lfloor K/2 \rfloor$ being the integer part of $K/2$. The gap between the two rates highlights the importance of binary decomposition that specifically takes advantage of the multi-mode structure in tensors.

The second lower bound is for any estimator $\tilde{\Theta}$ from the "unquantized" version of the continuous-valued tensor decomposition, which enables us to evaluate how much information is lost by quantizing a continuous-valued tensor to a binary one. Specifically, in Section 2.2, we show that our binary tensor model can be viewed as an entrywise quantization of a noisy continuous-valued tensor. We next consider an "unquantized" version of the model where the noisy entries, $\tilde{\mathcal{Y}} = \Theta + \mathcal{E}$, are observed directly without any quantization. We seek an estimator $\tilde{\Theta}$ by "denoising" the continuous-valued observation. The lower bound is obtained via an information-theoretical argument and is applicable to any estimator $\tilde{\Theta} \in \mathcal{D}(R, \alpha)$.

Theorem 3 (Minimax lower bound for continuous-valued tensors). *Suppose* $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ *is an order-*K *continuous-valued tensor from the model* $\mathcal{Y} = \Theta_{true} + \mathcal{E} \in \mathbb{R}^{d_1 \times \cdots d_K}$, *where* $\Theta_{true} \in \mathcal{D}(R, \alpha)$ *and* $\mathcal{E}$ *is a tensor of i.i.d. Gaussian entries. Suppose that* $R \leq \min_k d_k$ *and* $\max_k d_k \geq 8$. *Let* $\tilde{\Theta}$ *denote an estimator of* Θ_{true}. *Then there exist absolute constants* $\beta_0 \in (0,1)$ *and* $c_0 > 0$ *such that*

$$\inf_{\tilde{\Theta}} \sup_{\Theta_{true} \in \mathcal{D}(R,\alpha)} \mathbb{P}\left\{ Loss(\tilde{\Theta}, \Theta_{true}) \geq c_0 \min\left(\alpha, \sigma\sqrt{\frac{Rd_{\max}}{\prod_k d_k}}\right)\right\} \geq \beta_0. \tag{9}$$

The proof is given in the Appendix. This lower bound (9) quantifies the intrinsic hardness of the problem. In the next section, we explicitly evaluate the information loss of our tensor estimation method based on the quantized binary data, by comparing to the case when the latent tensor $\tilde{\mathcal{Y}}$ is fully observed without any quantization.

3.3 Phase diagram

The error bounds we have established depend on the signal bound α and the noise level σ. In this section, we define three regimes based on the signal-to-noise ratio, $\text{SNR} = \|\Theta\|_\infty/\sigma$, in which the tensor estimation exhibits different behaviors. We borrow the term "phase diagram" from chemistry to describe those different regimes, or phases. Table 1 and Figure 1 summarize the error bounds of the three phrases under the case when $d_1 = \ldots = d_K = d$. Our discussion focuses on the probit model, but similar patterns also hold for the logistic and Laplacian models.

The first phase is when the noise is weak, in that $\sigma \ll \alpha$ and $\text{SNR} \gg \mathcal{O}(1)$. In this regime, we see that the error bound in (7) scales as $\sigma \exp(\alpha^2/\sigma^2)$, suggesting that increasing the noise level would lead to an improved tensor estimation accuracy. This behavior may seem surprising. Such a phenomenon is not an artifact of the proof, but instead is intrinsic to 1-bit quantization. We also confirm this behavior in simulations in Section 5. In fact, when σ goes to zero, the problem essentially reverts to the noiseless case where an accurate estimation of Θ becomes impossible. To see this, we consider a simple example with a rank-1 signal tensor in the latent model (4). Two different coefficient tensors, $\Theta_1 = \boldsymbol{a}_1 \otimes \boldsymbol{a}_2 \otimes \boldsymbol{a}_3$ and $\Theta_2 = \text{sign}(\boldsymbol{a}_1) \otimes \text{sign}(\boldsymbol{a}_2) \otimes \text{sign}(\boldsymbol{a}_3)$, would lead to the same observation $\mathcal{Y}$ in the absence of noise, and thus recovery of Θ from $\mathcal{Y}$ becomes hopeless. Interestingly, adding a stochastic noise $\mathcal{E}$ to the signal tensor prior to 1-bit quantization completely changes the nature of the problem, and an efficient estimate can be obtained through the likelihood approach. In the 1-bit matrix/tensor completion literature, this phenomenon is referred to as "dithering" effect of random noise (Davenport et al., 2014).

Tensor type	$\text{SNR} \gg \mathcal{O}(1)$	$\mathcal{O}(1) \gtrsim \text{SNR} \gg \mathcal{O}(d^{-(K-1)/2})$	$\mathcal{O}(d^{-(K-1)/2}) \gtrsim \text{SNR}$
Binary	$\sigma e^{\alpha^2/\sigma^2} d^{-(K-1)/2}$	$\sigma d^{-(K-1)/2}$	α
Continuous	$\sigma d^{-(K-1)/2}$	$\sigma d^{-(K-1)/2}$	α

Table 1: Error rate for low-rank tensor estimation. For ease of presentation, we omit the constants in rate that depend on K and R.

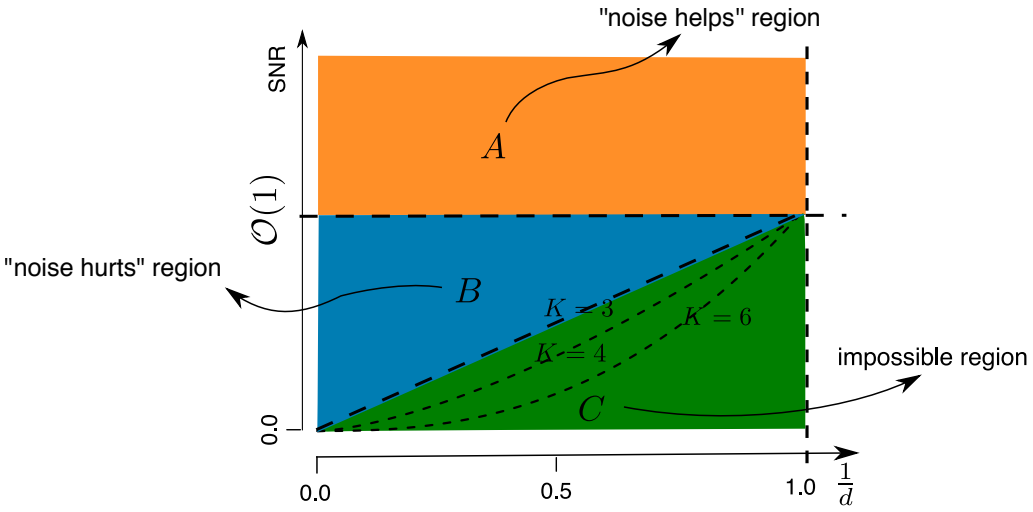


Figure 1: Phase diagram with respect to SNR. (A) "Noise helps" region (in yellow): the estimation error decreases with the noise . (B) "Noise hurts" region (in blue): the error increases with the noise. (C) Impossible region (in green): a consistent estimator of Θ is impossible. The dashed line between regions (B) and (C) depicts the boundary $d^{-(K-1)/2}$ as K varies. Note that the origin in the x-axis corresponds to the high-dimensional region, $d^{-(K-1)/2} \to 0$, that is of main interest.

The second phase is when the noise is comparable to the signal, in that $\mathcal{O}(1) \gtrsim \text{SNR} \gg \mathcal{O}(d^{-(K-1)/2})$. In this regime, the error bound in (7) scales linearly with σ. Comparing the lower bound (9) when estimating the signal from the unquantized tensor, to the upper bound (7) when estimating from a quantized one, the two error bounds match with each other. This suggests that 1-bit quantization induces very little loss of information towards the estimation of Θ in this regime. In other words, $\hat{\Theta}_{\text{MLE}}$, which is based on the quantized tensor, can achieve the same degree of accuracy as if it were given access to the completely unquantized measurements.

The third phase is when the noise completely dominates the signal, in that $\text{SNR} \lesssim \mathcal{O}(d^{-(K-1)/2})$. A consistent estimation of Θ becomes impossible. In this regime, a trivial zero estimator achieves the minimax rate.

4. Algorithm and Convergence Properties

4.1 Block relaxation algorithm

In this section, we first introduce an algorithm to solve (5), then study the algorithmic convergence. For notational convenience, we drop the subscript $\mathcal{Y}$ in $\mathcal{L}_{\mathcal{Y}}(\Theta)$ and simply write $\mathcal{L}(\Theta)$. The objective function $\mathcal{L}(\Theta)$ is concave in Θ when the link f is log-concave in $\theta_{i_1,\ldots,i_K}$. However, the feasible set $\mathcal{D}$ is nonconvex, and thus the optimization (5) is a non-convex problem. We utilize a formulation of the CP decomposition, and turn the optimization into a block-wise convex problem.

Specifically, write the mode-k factor matrices from (3) as

$$\boldsymbol{A}_k = \left\{\boldsymbol{a}_1^{(k)}, \ldots, \boldsymbol{a}_R^{(k)}\right\} \in \mathbb{R}^{d_k \times R}, k \in [K-1], \text{ and } \boldsymbol{A}_K = \left\{\lambda_1 \boldsymbol{a}_1^{(K)}, \ldots, \lambda_R \boldsymbol{a}_R^{(K)}\right\} \in \mathbb{R}^{d_K \times R}, \tag{10}$$

where, without loss of generality, we choose to rescale λ_k's into the last factor matrix. We denote by $\boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K)$ the collection of all block variables satisfying the above convention. Then the optimization problem (5) can be rewritten as

$$\max_{\boldsymbol{A}} \mathcal{L}\{\Theta(\boldsymbol{A})\}, \quad \text{subject to } \Theta(\boldsymbol{A}) \in \mathcal{D}. \tag{11}$$

Although the objective function in (11) is in general not concave in the K factor matrices jointly, it is concave in each factor matrix individually with all other factor matrices fixed. This feature enables a block relaxation type minimization, where we alternatively update one factor matrix at a time while keeping the others fixed. In each iteration, the update of each factor matrix involves solving a number of GLMs separately. To see this, let $\boldsymbol{A}_k^{(t)}$ denote the kth factor matrix at the tth iteration, and

$$\boldsymbol{A}_{-k}^{(t)} = \boldsymbol{A}_1^{(t+1)} \odot \cdots \odot \boldsymbol{A}_{k-1}^{(t+1)} \odot \boldsymbol{A}_{k+1}^{(t)} \odot \cdots \odot \boldsymbol{A}_K^{(t)}, \quad k = 1, \ldots, K.$$

Let $\mathcal{Y}(:, j(k), :)$ denote the sub-tensor of $\mathcal{Y}$ at the jth position of the kth mode, $j = 1, \ldots, d_k, k = 1, \ldots K$, and $\mathrm{vec}(\cdot)$ is the operator that turns a tensor into a vector. Then the update $\boldsymbol{A}_k^{(t+1)}$ can be obtained row-by-row by solving d_k separate GLMs, where each GLM takes $\mathrm{vec}\{\mathcal{Y}(:, j(k), :)\} \in \mathbb{R}^{(\prod_{i \neq k} d_i) \times 1}$ as the "response", $\boldsymbol{A}_{-k}^{(t)} \in \mathbb{R}^{(\prod_{i \neq k} d_i) \times R}$ as the "predictors", and the jth row of $\boldsymbol{A}_k$ as the "regression coefficient", $j = 1, \ldots, d_k, k = 1, \ldots K$. In each GLM, the effective number of predictors is R, and the effective sample size is $\prod_{i \neq k} d_i$. These separable, low-dimensional GLMs allow us to leverage the standard GLM solvers such as those in R/Python/MATLAB, as well as parallel processing that can further speed up the computation. After each iteration, we post-process the factor matrices $\boldsymbol{A}_k^{(t+1)}$ by performing a line search,

$$\gamma^* = \arg\max_{\gamma \in [0,1]} \mathcal{L}_{\mathcal{Y}} \left\{\gamma \boldsymbol{A}_k^{(t)} + (1-\gamma)\boldsymbol{A}_k^{(t+1)}\right\}, \quad \text{subject to } \|\Theta\|_\infty \leq \alpha.$$

We then update $\boldsymbol{A}_k^{(t+1)} = \gamma^* \boldsymbol{A}_k^{(t)} + (1-\gamma^*)\boldsymbol{A}_k^{(t+1)}$, and normalize the columns of $\boldsymbol{A}_k^{(t+1)}$. The full optimization procedure is summarized in Algorithm 1.

4.2 Algorithmic properties

Because the block relaxation algorithm monotonically increases the objective function, the convergence of the objective function is guaranteed whenever the $\mathcal{L}$ is bounded from above. We next study the convergence of the actual iterates $\boldsymbol{A}^{(t)}$ and $\Theta^{(t)} = \Theta\{\boldsymbol{A}^{(t)}\}$ resulting from Algorithm 1. To simplify the analysis, we set the hyper-parameter α to infinity, which essentially poses no prior on the tensor magnitude. We need the following assumptions.

(A1) (Regularity condition) The log-likelihood $\mathcal{L}(\boldsymbol{A})$ is continuous and the set $\{\boldsymbol{A}\colon \mathcal{L}(\boldsymbol{A}) \geq \mathcal{L}(\boldsymbol{A}^{(0)})\}$ is compact.

Algorithm 1 Binary tensor decomposition

Input: Binary tensor $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$, link function f, rank R, and entrywise bound α.
Output: Rank-R coefficient tensor Θ, along with the factor matrices $\boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K)$.

1: Initialize random matrices $\boldsymbol{A}^{(0)} = \left\{\boldsymbol{A}_1^{(0)}, \ldots, \boldsymbol{A}_K^{(0)}\right\}$ and iteration index $t = 0$.
2: **while** the relative increase in objective function $\mathcal{L}(\boldsymbol{A})$ is less than the tolerance **do**
3: Update iteration index $t \leftarrow t + 1$.
4: **for** $k = 1$ to K **do**
5: Obtain $\boldsymbol{A}_k^{(t+1)}$ by solving d_k separate GLMs with link function f.
6: **end for**
7: Line search to obtain γ^*.
8: Update $\boldsymbol{A}_k^{(t+1)} \leftarrow \gamma^* \boldsymbol{A}_k^{(t)} + (1-\gamma^*)\boldsymbol{A}_k^{(t+1)}$, for all $k \in [K]$.
9: Normalize the columns of $\boldsymbol{A}_k^{(t+1)}$ to be of unit-norm for all $k \leq K-1$, and absorb the scales into the columns of $\boldsymbol{A}_K^{(t+1)}$.
10: **end while**

(A2) (Strictly local maximum condition) Each block update in Algorithm 1 is well-defined; i.e., the GLM solution exists and is unique, and the corresponding sub-block in the Hession is non-singular at the solution.

(A3) (Local uniqueness condition) The set of stationary points of $\mathcal{L}(\boldsymbol{A})$ are isolated module scaling.

(A4) (Local Lipschitz condition) The tensor rank-R representation $\Theta = \Theta(\boldsymbol{A})$ is called locally Lipschitz at $\boldsymbol{A}^*$, if there exits two constants $c_1, c_2 > 0$ such that

$$c_1 \|\boldsymbol{A}' - \boldsymbol{A}''\|_F \leq \|\Theta(\boldsymbol{A}') - \Theta(\boldsymbol{A}'')\|_F \leq c_2 \|\boldsymbol{A}' - \boldsymbol{A}''\|_F,$$

for $\boldsymbol{A}', \boldsymbol{A}''$ sufficiently close to $\boldsymbol{A}^*$. Here $\boldsymbol{A}', \boldsymbol{A}''$ represent the block variables subject to convention (10).

All these are fairly mild conditions and are often imposed in the literature. Specifically, Assumption (A1) ensures that the maximum exists and the log-likelihood is bounded above. Therefore, the stopping rule of Algorithm 1 is well defined. Assumption (A2) asserts the negative-definiteness of the Hessian in the block coordinate $\boldsymbol{A}_k$. Note that the full Hession needs not to be negative-definite in all variables simultaneously. We consider this as a reasonably mild assumption, and similar conditions have often been imposed in numerous non-convex problems (Uschmajew, 2012; Zhou et al., 2013; Sun et al., 2017). Assumptions (A2)–(A4) guarantee the local uniqueness of the CP decomposition $\Theta = \Theta(\boldsymbol{A})$. The conditions exclude the case of rank-degeneracy; e.g., if the tensor Θ can be written in fewer than R factors, or if the columns of $\boldsymbol{A}_{-k}^{(t)}$ are linearly dependent in the GLM update. They also exclude the case of non-unique decompositions; e.g., if the decomposition Θ can be smoothly changed beyond scaling of the factors. These conditions are fairly mild for tensors of order 3 or higher. For more discussion on decomposition uniqueness and its implication in the optimization, we refer to Kruskal (1977); Uschmajew (2012); Zhou et al. (2013).

Proposition 1 (Algorithmic convergence). *Suppose Assumptions (A1)-(A3) hold.*

(i) (Global convergence) Any sequence $\boldsymbol{A}^{(t)} = \left\{\boldsymbol{A}_1^{(t)}, \ldots, \boldsymbol{A}_K^{(t)}\right\}$ *generated by Algorithm 1 converges to a stationary point of* $\mathcal{L}(\boldsymbol{A})$.

(ii) (Locally linear convergence) Let $\boldsymbol{A}^*$ *be a local maximizer of* $\mathcal{L}$. *There exists an* ε*-neighborhood of* $\boldsymbol{A}^*$, *such that, for any staring point* $\boldsymbol{A}^{(0)}$ *in this neighborhood, the iterates* $\boldsymbol{A}^{(t)}$ *of Algorithm 1 linearly converge to* $\boldsymbol{A}^*$,

$$\|\boldsymbol{A}^{(t)} - \boldsymbol{A}^*\|_F \leq \rho^t \|\boldsymbol{A}^{(0)} - \boldsymbol{A}^*\|_F,$$

where $\rho \in (0,1)$ *is a contraction parameter. Furthermore, if Assumption (A4) holds at* $\boldsymbol{A}^*$, *then there exists a constant* $C > 0$ *such that*

$$\|\Theta(\boldsymbol{A}^{(t)}) - \Theta(\boldsymbol{A}^*)\|_F \leq C\rho^t \|\Theta(\boldsymbol{A}^{(0)}) - \Theta(\boldsymbol{A}^*)\|_F.$$

Proposition 1(ii) shows that every local maximizer of $\mathcal{L}$ is an attractor of Algorithm 1. This is a nice property that ensures the exponential decay of the error near a local maximum. Moreover, it reflects some intrinsic difference between tensor decomposition and matrix decomposition, in that the same property often fails for the matrix case. Consider an example of a 2-by-2 matrix. Suppose that the local maximizer is $\Theta^* = \Theta^*(\boldsymbol{e}_1, \boldsymbol{e}_2) = \boldsymbol{e}_1^{\otimes 2} + \boldsymbol{e}_2^{\otimes 2}$, where $\boldsymbol{e}_1, \boldsymbol{e}_2$ are canonical vectors in $\mathbb{R}^2$. There is no region of attraction near the block variable $\boldsymbol{A}^* = (\boldsymbol{e}_1, \boldsymbol{e}_2)$. In fact, one can construct a point $\boldsymbol{A}^{(0)} = (\boldsymbol{a}_1, \boldsymbol{a}_2)$, with $\boldsymbol{a}_1 = (\sin\theta, \cos\theta)'$, and $\boldsymbol{a}_2 = (\cos\theta, -\sin\theta)'$. The point $\boldsymbol{A}^{(0)}$ can be made arbitrarily close to $\boldsymbol{A}^*$ by tuning θ, but the algorithm iteration initialized from $\boldsymbol{A}^{(0)}$ would never converge to $\boldsymbol{A}^*$. This behavior occurs when the matrix has degenerate singular values. In contract, a 2-by-2-by-2 tensor problem with the maximizer $\tilde{\Theta}^* = \tilde{\Theta}^*(\boldsymbol{e}_1, \boldsymbol{e}_2) = \boldsymbol{e}_1^{\otimes 3} + \boldsymbol{e}_2^{\otimes 3}$ possesses locally unique decomposition, and thus isolated stationary points. The block variable $\boldsymbol{A}^*$ exhibits a basin of attraction with respect to the tensor objective.

Combining Proposition 1 and Theorem 1, we have the following theorem.

Theorem 4 (Empirical performance). *Suppose* $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$ *is a binary tensor under the Bernoulli tensor model (2) with parameter* $\Theta_{true} = \Theta(\boldsymbol{A}_{true})$. *Let* $\boldsymbol{A}^{(t)}$ *denote a sequence of estimators generated from Algorithm 1, with the initial point* $\boldsymbol{A}^{(0)}$ *and the limiting point* $\boldsymbol{A}^*$. *Suppose that the initialization error* $Loss(\Theta(\boldsymbol{A}^{(0)}), \Theta_{true})$ *is bounded by some constant* $\varepsilon > 0$, *and that* $\boldsymbol{A}^*$ *satisfies* $\mathcal{L}[\Theta(\boldsymbol{A}^*)] \geq \mathcal{L}(\Theta_{true})$. *Suppose Assumptions (A1)-(A4) hold. Then we have, with probability at least* $1 - exp(-C' \log K \sum_k d_k)$,

$$Loss\left(\Theta(\boldsymbol{A}^{(t)}), \Theta_{true}\right) \leq \underbrace{C_1 \rho^t \varepsilon}_{\text{algorithmic error}} + \underbrace{\frac{C_2 L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}}}_{\text{statistical error}}, \tag{12}$$

where $\rho \in (0,1)$ *is a contraction parameter, and* $C_1, C_2 > 0$ *are two constants.*

Theorem 4 provides the estimation error of the actual estimator from our Algorithm 1 at each iteration. The bound (12) consists of two terms: the first term is the computational error, and the second is the statistical error. The computational error decays exponentially with the number of iterations, whereas the statistical error remains the same as t grows. The statistical error is unavoidable, as it reflects the intrinsic error due to estimating from

a noisy tensor; see also Theorem 2. The bound (12) thus reveals the interplay between the computational and the statistical errors. Moreover, for tensors with $d_1 = \ldots = d_K = d$, when the iteration number satisfies that,

$$t \geq T = \log_{1/\rho} \left(\frac{C_1 \varepsilon}{\frac{C_2 L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}}} \right) \asymp \log_{1/\rho} \left\{ d^{(k-1)/2} \right\},$$

the computational error is to be dominated by the statistical error.

4.3 Missing data, rank selection and computational complexity

When some tensor entries $y_{i_1,\ldots,i_K}$ are missing, we replace the objective function $\mathcal{L}_{\mathcal{Y}}(\Theta)$ with $\sum_{(i_1,\ldots,i_K)\in\Omega} \log f(q_{i_1,\ldots,i_K}\theta_{i_1,\ldots,i_K})$, where $\Omega \subset [d_1] \times \cdots \times [d_K]$ is the index set for non-missing entries. That is, we model the observed entries only, and exclude the missing entries in the model fitting. This same strategy to handle missing values has been used for continuous-valued tensor decomposition (Acar et al., 2010). For implementation, we modify line 5 in Algorithm 1, by fitting GLMs to the data for which $y_{i_1,\ldots,i_K}$ are observed. Other steps in Algorithm 1 are amendable to missing data accordingly. This approach requires that there are no completely missing sub-tensors $\mathcal{Y}(:, j(k), :)$, which is a fairly mild condition. This is similar to the coherence condition in the matrix completion problem; for instance, if an entire row or column of a matrix is missing, it is impossible to recover its true decomposition.

As a by-product, our tensor decomposition output can also be used for missing value prediction. That is, we predict the missing values $Y_{i_1,\ldots,i_K}$ using $f(\hat{\Theta}_{i_1,\ldots,i_K})$, where $\hat{\Theta}$ is the coefficient tensor estimated from the observed entries. Note that the predicted values are always between 0 and 1, which can be interpreted as a prediction for $\mathbb{P}(Y_{i_1,\ldots,i_K} = 1)$.

Algorithm 1 requires the rank of Θ as an input. Estimating an appropriate rank given the data is thus of practical importance. We adopt the usual Bayesian information criterion (BIC), and choose the rank that minimizes BIC; i.e.,

$$\hat{R} = \arg\min_{R\in\mathbb{R}_+} \mathrm{BIC}(R) = \arg\min_{R\in\mathbb{R}_+} \left[-2\mathcal{L}_{\mathcal{Y}}\{\hat{\Theta}(R)\} + p_e(R) \log \left(\prod_k d_k \right) \right],$$

where $\hat{\Theta}(R)$ is the estimated coefficient tensor Θ under the working rank R, and $p_e(R)$ is the effective number of parameters. This criterion aims to balance between the goodness-of-fit for the data and the degree of freedom in the population model. The empirical performance of this criterion is investigated in Section 5.

Finally, the computational complexity of our algorithm is $O(R^3 \prod_k d_k)$ for each iteration. The per-iteration computational cost scales linearly with the tensor dimension, and this complexity matches the classical continuous-valued tensor decomposition. More precisely, the update of $\boldsymbol{A}_k$ involves solving d_k separate GLMs. Solving those GLMs requires $O(R^3 d_k + R^2 \prod_k d_k)$, and therefore the cost for updating K factors in total is $O(R^3 \sum_k d_k + R^2 K \prod_k d_k)$. We further report the computation time in Section 5.

5. Simulations

5.1 CP tensor model

In this section, we first investigate the finite-sample performance of our method when the data indeed follows the CP tensor model. Specifically, we consider an order-3 dimension-(d, d, d) binary tensor $\mathcal{Y}$ generated from the threshold model (4), where $\Theta_{\text{true}} = \sum_{r=1}^{R} \boldsymbol{a}_r^{(1)} \times \boldsymbol{a}_r^{(2)} \times \boldsymbol{a}_r^{(3)}$, and the entries of $\boldsymbol{a}_r^k$ are i.i.d. drawn from Uniform$[-1, 1]$. Without loss of generality, we scale Θ_{true} such that $\|\Theta_{\text{true}}\|_\infty = 1$. The binary tensor $\mathcal{Y}$ is generated based on the entrywise quantization of the latent tensor $(\Theta^{\text{true}} + \mathcal{E})$, where $\mathcal{E}$ consists of i.i.d. Gaussian entries. We vary the rank $R \in \{1, 3, 5\}$, the tensor dimension $d \in \{20, 30, \ldots, 60\}$, and the noise level $\sigma \in \{10^{-3}, 10^{-2.5}, \ldots, 10^{0.5}\}$. We use BIC to select the rank, use the logistic link function, and report the estimation error averaged across $n_{\text{sim}} = 30$ replications.

Figure 2(a) plots the estimation error $\text{Loss}(\Theta_{\text{true}}, \hat{\Theta}_{\text{MLE}})$ as a function of the tensor dimension d while holding the noise level fixed at $\sigma = 10^{-0.5}$ for three different ranks $R \in \{1, 3, 5\}$. It is seen that the estimation error of the constrained MLE decreases as the dimension increases. Consistent with our theoretical results, the decay in the error appears to behave on the order of d^{-1}. A higher-rank tensor tends to yield a larger recovery error, as reflected by the upward shift of the curves as R increases. Indeed, a higher rank means a higher intrinsic dimension of the problem, thus increasing the difficulty of the estimation.

Figure 2(b) plots the estimation error as a function of the noise level σ while holding the dimension fixed at $d = 50$ for three different ranks $R \in \{1, 3, 5\}$. A larger estimation error is observed when the noise is either too small or too large. The non-monotonic behavior confirms the phase transition with respect to the SNR. Particularly, in the high SNR regime, the random noise is seen to improve the recovery accuracy. This is consistent to our theoretical result on the "dithering" effects brought by stochastic noise.

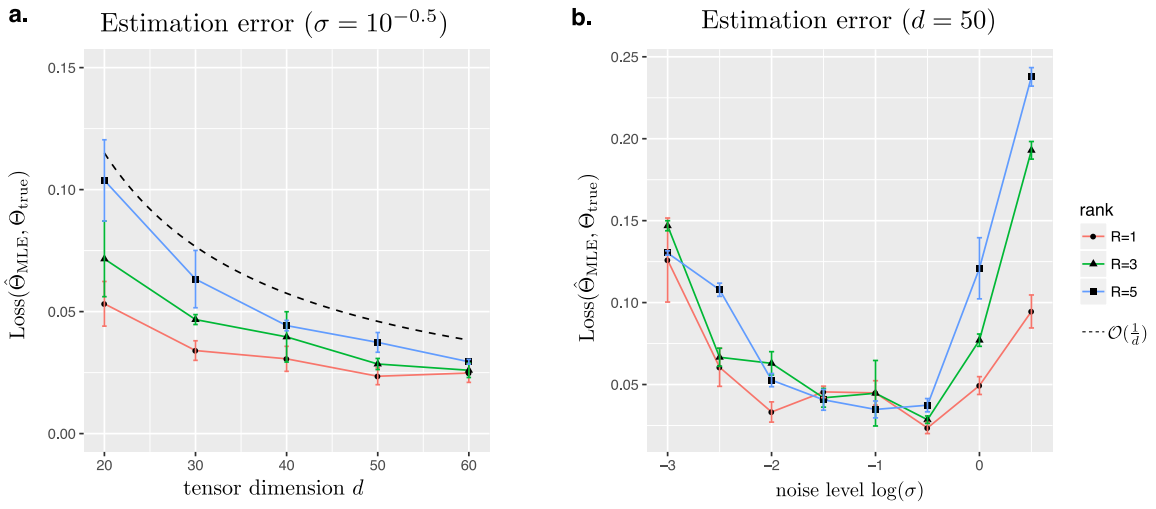


Figure 2: Estimation error of binary tensor decomposition. (a) Estimation error as a function of the tensor dimension $d = d_1 = d_2 = d_3$. (b) Estimation error as a function of the noise level.

	$\sigma = 0.1$			$\sigma = 0.01$		
True rank	$d = 20$	$d = 40$	$d = 60$	$d = 20$	$d = 40$	$d = 60$
$R = 5$	4.9 (0.2)	5 (0)	5 (0)	4.8 (1.0)	5 (0)	5 (0)
$R = 10$	8.7 (0.9)	10 (0)	10 (0)	8.8 (0.4)	10 (0)	10 (0)
$R = 20$	17.7(1.7)	20.4(0.5)	20.2(0.5)	16.4(0.5)	20.4(0.5)	20.6(0.5)
$R = 40$	36.8(1.1)	39.6(1.7)	40.2(0.4)	36.0(1.2)	38.8(1.6)	40.3(1.1)

Table 2: Rank selection in binary tensor decomposition via BIC. The selected rank is averaged across 30 simulations, with the standard error shown in the parenthesis.

We next assess the tensor rank selection by BIC. We consider the tensor dimension $d \in \{20, 40, 60\}$ and rank $R \in \{5, 10, 20, 40\}$. This way, in some of the combinations, the rank equals or exceeds the tensor dimension. We set the noise level $\sigma \in \{0.1, 0.01\}$ such that the noise is neither negligible nor overwhelming. For each combination, we simulate the tensor data following the Bernoulli tensor model (2). We minimize BIC using a grid search from $R - 5$ to $R + 5$. Table 2 reports the selected rank averaged over $n_{\text{sim}} = 30$ replications, with the standard error shown in the parenthesis. It is seen that, when $d = 20$, the selected rank is slightly smaller than the true rank, whereas for $d \geq 40$, the selection is accurate. This agrees with our expectation, as in tensor decomposition, the total number of entries corresponds to the sample size. A larger d implies a larger sample size, so the BIC selection becomes more accurate.

We also evaluate the numerical stability of our optimization algorithm. Although Algorithm 1 has no theoretical guarantee to land to the global optimum, in practice, we often find that the converged points are satisfactory, in that the corresponding objective values are close to the objective function evaluated at the true parameter, $\mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}})$. As an illustration, Figure 3 shows the typical trajectories of the objective function under different tensor dimensions and ranks. The dashed line is the objective value at the true parameter, $\mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}})$. It is seen that the algorithm generally converges quickly upon random initializations, usually taking fewer than 8 iterations for the relative change in the objective to be below 3%, even for a large d and R. The average computation time per iteration is shown in the plot legend. For instance, when $d = 60$ and $R = 10$, each iteration of Algorithm 1 on average takes fewer than 3 seconds.

5.2 Stochastic multi-way block model

We next evaluate our method under the stochastic multi-way block model, which can be viewed as a higher-order generalization of the stochastic block model that is commonly used for random graphs, network analysis, and community detection (Anandkumar et al., 2014; Abbe, 2017). Under this model, the signal tensor does not necessarily admit a low-rank structure. Specifically, we generate $\mathcal{Y}$ of dimension $d = d_1 = d_2 = d_3$, where we vary $d \in \{20, 30, 40, 50, 60\}$. The entries in $\mathcal{Y}$ are realizations of independent Bernoulli variables with the probability tensor Θ, which is partitioned into five blocks along each of the modes,

$$\text{Probit}^{-1}(\Theta) = \mathcal{C} \times_1 \boldsymbol{N}_1 \times_2 \boldsymbol{N}_2 \times_3 \boldsymbol{N}_3.$$

Here $\mathcal{C} = [\![c_{m_1 m_2 m_3}]\!] \in \mathbb{R}^{5\times 5\times 5}$ is a core tensor corresponding to the block-means in a probit scale, $\boldsymbol{N}_1, \boldsymbol{N}_2, \boldsymbol{N}_3 \in \{0, 1\}^{5\times d}$ are membership matrices indicating the block allocation along

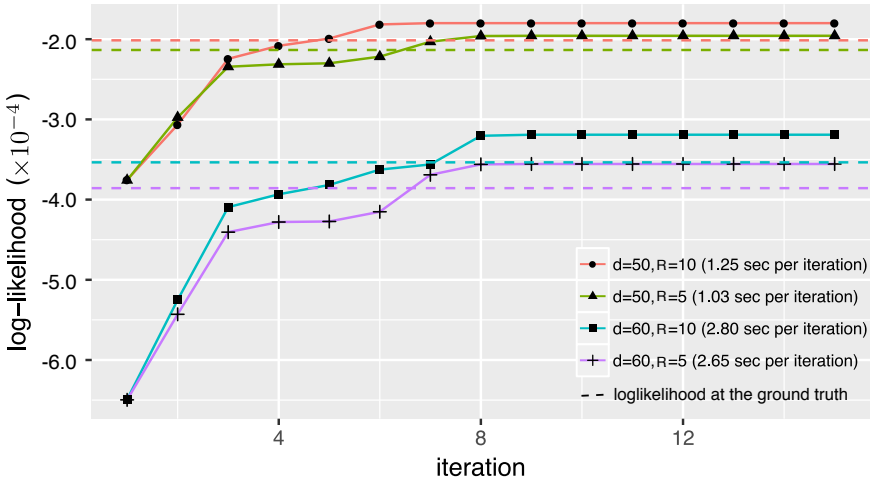


Figure 3: Trajectory of the objective function over iterations with varying d and R.

each of the mode, and $m_1, m_2, m_3 = 1, \ldots, 5$ are block indices. We first generate the block means $\{c_{m_1 m_2 m_3}\}$ in the following ways:

- Combinatorial-mean model: $c_{m_1 m_2 m_3} \overset{\text{i.i.d.}}{\sim} \text{Uniform}[-1, 1]$, i.e., each three-way block has its own mean, independent of each other.
- Additive-mean model: $c_{m_1 m_2 m_3} = c^1_{m_1} + \mu^2_{m_2} + \mu^3_{m_3}$, where $\mu^1_{m_1}$, $\mu^2_{m_2}$ and $\mu^3_{m_3}$ are i.i.d. drawn from Unif$[-1, 1]$.
- Multiplicative-mean model: $c_{m_1 m_2 m_3} = c^1_{m_1} \mu^2_{m_2} \mu^3_{m_3}$, and the rest of setup is the same as the additive-mean model.

We evaluate our method in terms of the accuracy of recovering the latent tensor Θ given the binary observations. Table 3 reports the relative loss, the estimated rank, and the running time, averaged over $n_{\text{sim}} = 30$ data replications, for the above three sub-models. The relative loss is computed as $\|\hat{\Theta}_{\text{MLE}} - \Theta_{\text{true}}\|_F / \|\Theta_{\text{true}}\|_F$. It is seen that our method is able to recover the signal tensors well in all three scenarios. As an illustration, we also plot one typical realization of the true signal tensor, the input binary tensor, and the recovered signal tensor for each sub-model in Table 3. It is interesting to see that, not only the block structure but also the tensor magnitude are well recovered. We also remark that, there are two potential model misspecifications in this example. First, the additive and combinatorial-mean models do not follow an exact CP structure. Second, the data has been generated from a probit model, but we always fit with a logistic link. Our method is shown to maintain a reasonable performance under potential model misspecification.

5.3 Comparison with alternative methods

We next compare our method with a number of alternative solutions for binary tensor factorization (BTF).

- Boolean tensor factorization (BooleanTF) (Miettinen, 2011; Erdos and Miettinen, 2013b; Rukat et al., 2018). This method decomposes a binary tensor into binary

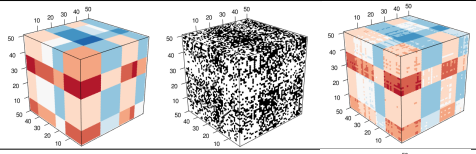
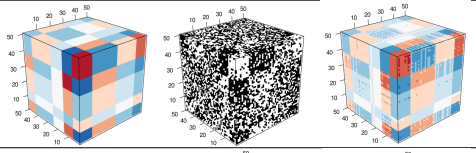
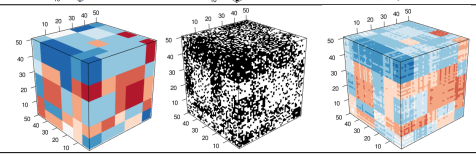
Block model	Experiment			Relative Loss	Rank Estimate	Time (sec)
	True signal	Input tensor	Output tensor			
Additive				0.23(0.05)	1.9(0.3)	4.23(1.62)
Multiplicative				0.22(0.07)	1.0(0.0)	1.70(0.09)
Combinatorial				0.48(0.04)	6.0(0.9)	10.4(3.4)

Table 3: Latent tensor recovery. Columns 2–4 are color images of the simulated tensor under different block mean models. Reported are the relative loss, the estimated rank, and the running time, averaged over 30 data replications, with the standard error shown in the parenthesis.

factors, then recovers the binary entries based on a set of logic rules among the factors. We use the implementation of Rukat et al. (2018).

- Bayesian tensor factorization (BTF_Bayeisan) (Rai et al., 2014). This method uses expectation-maximization to decompose a binary tensor into continuous-valued factors. It imposes a Gaussian prior on the factor entries, and a multiplicative gamma process prior on the factor weights $\{\lambda_r\}$.

- Bernoulli tensor factorization with gradient descent (BTF_Gradient) (Hong et al., 2018). This method uses a gradient descent algorithm to decompose a binary tensor into continuous-valued factors. We use the implementation in the toolbox of Matlab.

For easy reference, we denote our method by BTF_Alternating, and our implementation can be found at `https://github.com/Miaoyanwang/Binary-Tensor`. These four methods differ in several ways. BooleanTF is different from the other three in both the cost function and the output format. The rest are all based on the Bernoulli model (2), but with different implementations. BTF_Bayesian employs a Bayesian approach, whereas the other two are frequentist solutions. BTF_Gradient and our method, BTF_Alternating, share the same model, but utilize different optimization algorithms. So the two methods complement each other. On the other hand, in this article, we provide not only the algorithm-specific convergence properties, but also algorithm-independent statistical properties including the statistical convergence rate, SNR phase diagram, and mini-max rate. These results are not available in Hong et al. (2018) who proposed BTF_Gradient, but these properties hold for BTF_Gradient as well.

We apply the four methods with the default parameters, while selecting the rank R using the recommended approach of each. For our method BTF_Alternating, we use the proposed BIC to select the rank. Since BTF_Gradient does not provide any rank selection criterion, we apply the same R selected by our BIC. For BTF_Alternating, we also set the hyper-parameter α to infinity, which essentially poses no prior on the tensor magnitude. Besides, because BTF_Bayesian only supports the logistic link, we use the logistic link in all three BTF methods.

We evaluate each method by two metrics. The first metric is the root mean square error, RMSE $= \left(\sqrt{\prod_k d_k}\right)^{-1} \|\widehat{\mathbb{E}(\mathcal{Y})} - \mathbb{E}(\mathcal{Y})\|_F$, where $\widehat{\mathbb{E}(\mathcal{Y})}$ denotes the estimated success probability tensor. For BooleanTF, this quantity is represented as the posterior mean of $\mathcal{Y}$ (Miettinen, 2011), and for the other three methods, $\widehat{\mathbb{E}(\mathcal{Y})} = \text{logit}(\hat{\Theta})$. The second metic is the misclassification error rate, MER $= \left(\prod_k d_k\right)^{-1} \|\mathbb{1}_{\widehat{\mathbb{E}(\mathcal{Y})} \geq 0.5} - \mathbb{1}_{\mathbb{E}(\mathcal{Y}) \geq 0.5}\|_0$. Here the indicator function is applied to tensors in an element-wise manner, and $\|\cdot\|_0$ counts the number of non-zero entries in the tensor. These metrics reflect two aspects of the construction error. RMSE summarizes the total estimation error in the success probabilities, whereas MER summarizes the classification errors among 0's and 1's.

We simulate data from two different models, and in both cases, the signal tensors do not necessarily follow an exact low-rank CP structure. Therefore, in addition to method comparison, it also allows us to evaluate the robustness of our method under potential model misspecification.

The first model is a logic boolean tensor model following the setup in Rukat et al. (2018). We first simulate noiseless tensors $\mathcal{Y} = [\![y_{ijk}]\!]$ from the following model,

$$y_{ijk} = \bigvee_{r=1}^{R} \bigwedge_{ijk} a_{ir} b_{jr} c_{kr}, \text{ with } a_{ir} \sim \text{Ber}(p^a_{ir}),\ b_{jr} \sim \text{Bernoulli}(p^b_{jr}), c_{kr} \sim \text{Bernoulli}(p^c_{kr}),$$

where the binary factor entries $\{a_{ir}\}$, $\{b_{jr}\}$, $\{c_{kr}\}$ are mutually independent with each other, the factor probabilities $\{p^a_{ir}\}$, $\{p^b_{jr}\}$, $\{p^c_{kr}\}$ are generated i.i.d. from Beta(2,4), and $\vee$ and $\wedge$ denote the logical OR and AND operations, respectively. Equivalently, the tensor entry is 1 if and only if there exists one or more components in which all corresponding factor entries are 1. It is easy to verify that

$$\mathbb{E}(y_{ijk} | \{p^a_{ir}, p^b_{jr}, p^c_{kr}\}) = 1 - \prod_{r=1}^{R} \left(1 - p^a_{ir} p^b_{jr} p^c_{kr}\right).$$

We then add contamination bias to $\mathcal{Y}$ by flipping the tensor entries $0 \leftrightarrow 1$ i.i.d. with probability 0.1. We consider the tensor dimension $d_1 = d_2 = d_3 = 50$, and the boolean rank $R \in \{10, 15, 20, 25, 30\}$.

Figure 4(a)-(b) shows the performance comparison based on $n_{\text{sim}} = 30$ replications. It is seen that the three BTF methods outperform BooleanTF in RMSE. This shows the advantage of a probabilistic model, upon which all three BTF methods are built. In contrast, BooleanTF seeks patterns in a specific data, but does not address population estimation. For classification, BooleanTF performs reasonably well in distinguishing 0's versus 1's, which agrees with the data mining nature of BooleanTF. It is also interesting to see that MER peaks at $R = 20$. Further investigation reveals that this setting corresponds to the case when

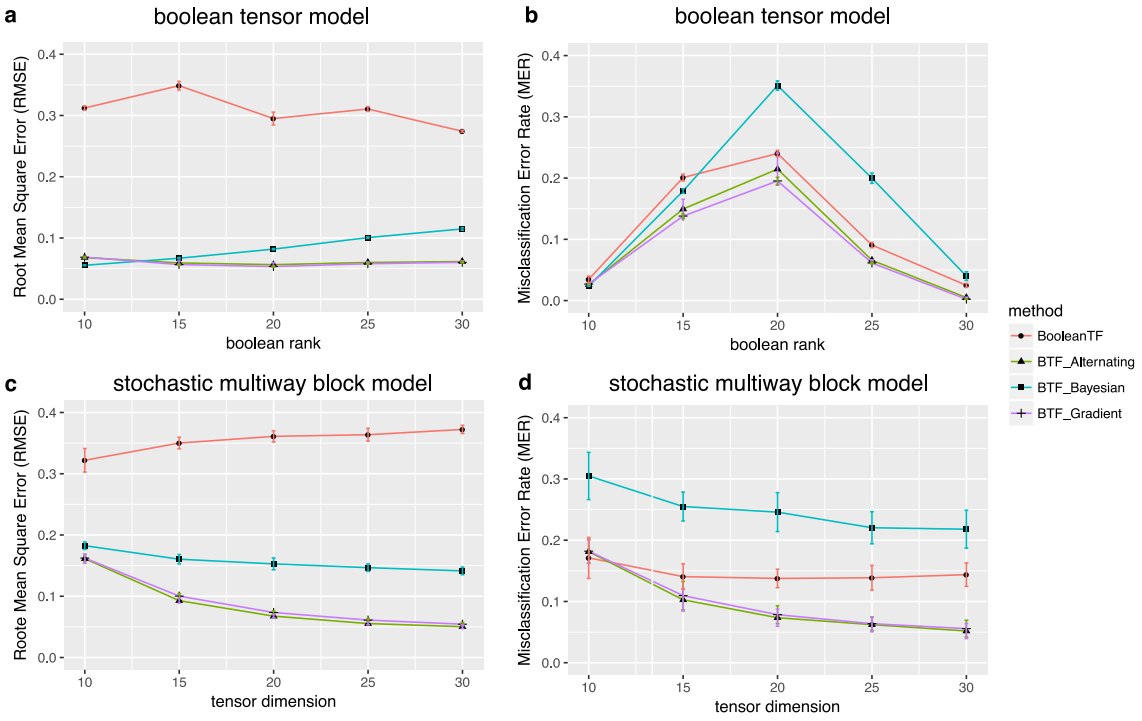


Figure 4: Performance comparison in terms of root mean squared error and misclassification error rate. (a)-(b) Estimation errors for the boolean tensor model. (c)-(d) Estimation errors for the stochastic multiway block model.

the Bernoulli probabilities $\mathbb{E}(\mathcal{Y})$ concentrate around 0.5, which becomes particularly challenging for classification. Actually, the average Bernoulli probability for R =10, 15, 20, 25, 30 is 0.31, 0.44, 0.53, 0.61, 0.68, respectively. Figure 4(b) also shows that BTF_Alternating and BTF_Gradient achieve a smaller classification error than BTF_Bayesian. One possible explanation is that the normal prior in BTF_Bayesian has a poor distinguishing power around $\theta \approx 0$, which corresponds to the hardest case when Bernoulli probability ≈ 0.5.

The second model is the stochastic multi-way block model considered in Section 5.2, with the block means $\{c_{m_1 m_2 m_3}\}$ generated from the combinatorial-mean sub-model. Figure 4(c)-(d) shows the performance comparison, and a similar pattern is observed. The two frequentist-type BTF methods, BTF_Gradient and BTF_Alternating, behave numerically similarly, and they outperform the other alternatives. In particular, the BTF methods exhibit decaying estimation errors, whereas BooleanTF appears to flatten out as dimension grows. This observation suggests that, compared to the algorithmic error, the statistical error is likely more dominating in this setting.

6. Real-world Data Applications

We next illustrate our binary tensor decomposition method using a number of real-world datasets of binary tensors, with applications ranging from social networks, email communication networks, to brain structural connectivities. We consider two tasks: one is tensor completion, and the other is clustering along one of the tensor modes, both of which are based upon the proposed binary tensor decomposition.

The datasets include:

- *Kinship* (Nickel et al., 2011): This is a $104 \times 104 \times 26$ binary tensor consisting of 26 types of relations among a set of 104 individuals in Australian Alyawarra tribe. The data was first collected by Denham and White (2005) to study the kinship system in the Alyawarra language. The tensor entry $\mathcal{Y}(i, j, k)$ is 1 if individual i used the kinship term k to refer to individual j, and 0 otherwise.

- *Nations* (Nickel et al., 2011): This is a $14 \times 14 \times 56$ binary tensor consisting of 56 political relations of 14 countries between 1950 and 1965. The tensor entry indicates the presence or absence of a political action, such as "treaties", "sends tourists to", between the nations. We note that the relationship between a nation and itself is not well defined, so we exclude the diagonal elements $\mathcal{Y}(i, i, k)$ from the analysis.

- *Enron* (Zhe et al., 2016): This is a $581 \times 124 \times 48$ binary tensor consisting of the three-way relationship, (sender, receiver, time), from the Enron email dataset. The Enron data is a large collection of emails from Enron employees that covers a period of 3.5 years. Following Zhe et al. (2016), we take a subset of the Enron data and organize it into a binary tensor, with entry $\mathcal{Y}(i, j, k)$ indicating the presence of emails from a sender i to a receiver j at a time period k.

- *HCP* (Wang et al., 2017a): This is a $68 \times 68 \times 212$ binary tensor consisting of structural connectivity patterns among 68 brain regions for 212 individuals from Human Connectome Project (HCP). All the individual images were preprocessed following a standard pipeline (Zhang et al., 2018), and the brain was parcellated to 68 regions-of-interest following the Desikan atlas (Desikan et al., 2006). The tensor entries encode the presence or absence of fiber connections between those 68 brain regions for each of the 212 individuals.

The first task is binary tensor completion, where we apply tensor decomposition to predict the missing entries in the tensor. We compare our binary tensor decomposition method using a logistic link function with the classical continuous-valued tensor decomposition method. Specifically, we split the tensor entries into 80% training set and 20 % testing set, while ensuring that the nonzero entries are split the same way between the training and testing data. The entries in the testing data are masked as missing and then predicted based on the tensor decomposition from the training data. Table 4 reports the area under the ROC curve (AUC) and RMSE, averaged over five random splits of the training and testing data. It is clearly seen that the binary tensor decomposition substantially outperforms the classical continuous-valued tensor decomposition. In all datasets, the former obtains a much higher AUC and mostly a lower RMSE. We also report in Table 4 the percentage of

Dataset	Non-zeros	Tensor decomposition method			
		Binary (logistic link)		Continuous-valued	
		AUC	RMSE	AUC	RMSE
Kinship	3.80%	0.9708	1.2×10^{-4}	0.9436	1.4×10^{-3}
Nations	21.1%	0.9169	1.1×10^{-2}	0.8619	2.2×10^{-2}
Enron	0.01%	0.9432	6.4×10^{-3}	0.7956	6.3×10^{-5}
HCP	35.3%	0.9860	1.3×10^{-3}	0.9314	1.4×10^{-2}

Table 4: Tensor completion for the four real-world binary tensor datasets. Two methods are compared: the proposed binary tensor decomposition, and the classical continuous-valued tensor decomposition.

nonzero entries for each data. It is seen that our decomposition method performs well even in the sparse setting. For instance, for the Enron dataset, only 0.01% of the entries are non-zero. The classical decomposition almost blindly assigns 0 to all the hold-out testing entires, resulting in a poor AUC of 79.6%. By comparison, our binary tensor decomposition achieves a much higher classification accuracy, with AUC = 94.3%.

The second task is clustering. We have carried out the clustering analysis on two datasets, *Nations* and *HCP*. We did not apply to the other two, *Kinship* and *Enron*, because there is no annotation information for the individuals in those two datasets. For the *Nations* dataset, we utilize a two-step procedure by first applying the proposed binary tensor decomposition method with the logistic link, then applying the K-means clustering along the country mode from the decomposition. In the first step, the BIC criterion suggests $R = 9$ factors, and in the second step, the classical elbow method selects 5 clusters out of the 9 components. Figure 5(a) plots the 9 tensor factors along the country mode. It is interesting to observe that the countries have been partitioned into one group containing those from the communist bloc, two groups from the western bloc, two groups from the neutral bloc, and Brazil forming its own group. To gain further insight, we also plot the top four relation types based on their loadings in the tensor factors along the relationship mode in Figure 5(b). The partition of the countries is seen to be consistent with their relationship patterns in the adjacency matrices. Indeed, those countries belonging to the same group tend to have similar linking patterns with other countries, as reflected by the block structure in Figure 5(b).

We also perform the clustering analysis on the dataset *HCP*. Again we apply the decomposition method with the logistic link first, and the BIC suggests the rank $R = 6$. Figure 6 plots the heatmap for the top 6 tensor components across the 68 brain regions, and Figure 7 shows the edges with high loadings based on the tensor components. Edges are overlaid on the brain template BrainMesh_ICBM152 (Xia et al., 2013), and nodes are color coded based on their regions. We see that the brain regions are spatially separated into several groups and that the nodes within each group are more densely connected with each other. We also observe some interesting spatial patterns in the brain connectivity. For instance, the edges captured by tensor component 2 are located within the cerebral hemisphere. The detected edges are association tracts consisting of the long association fibers, which connect different lobes of a hemisphere, and the short association fibers, which connect different gyri within a single lobe. In contrast, the edges captured by tensor component 3 are lo-

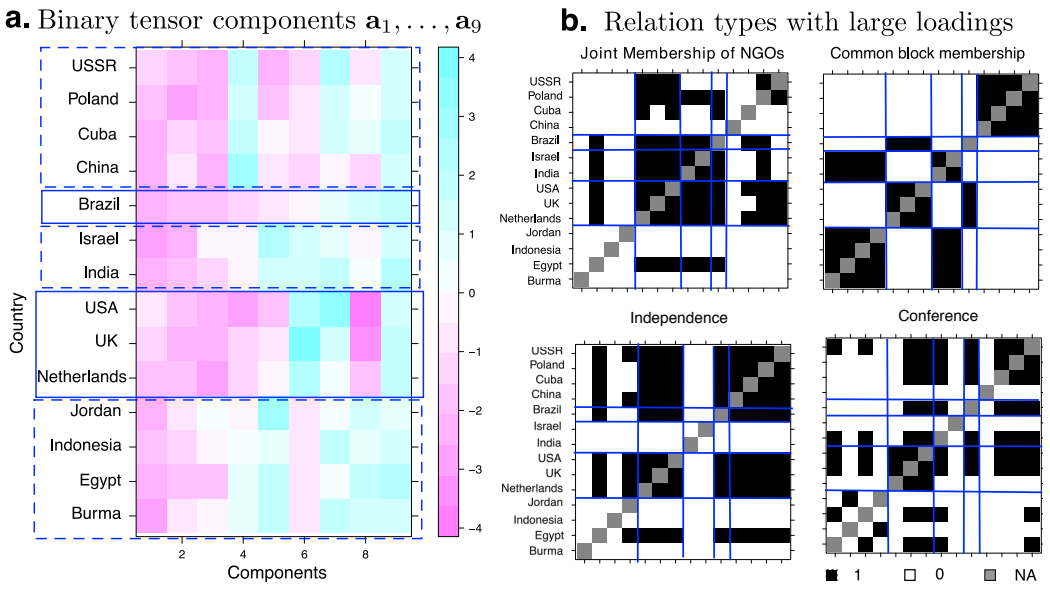


Figure 5: Analysis of the *Nations* dataset. (a) Binary tensor components. Top 9 tensor components in the country mode from the binary tensor decomposition. The overlaid box depicts the results from the K-means clustering. (b) Relation types with large loadings. Top four relationships identified from the top tensor components in the relation mode.

cated across the two hemispheres. Among the nodes with high connection intensity, we identify superior frontal gyrus, which is known to be involved in self-awareness and sensory system (Goldberg et al., 2006). We also identify corpus callosum, which is known as the largest commissural tract in the brain that connects two hemispheres. This is consistent with brain anatomy that suggests the key role of corpus callosum in facilitating interhemispheric connectivity (Roland et al., 2017). Moreover, the edges shown in tensor component 4 are mostly located within the frontal lobe, whereas the edges in component 5 connect the frontal lobe with parietal lobe.

7. Conclusion

Many real-world tensors consist of binary observations. In this article, we have presented a new binary tensor decomposition method. Under suitable link functions, we have shown that the unknown parameter tensor can be accurately and efficiently recovered. When the infinity norm of the unknown tensor is bounded by a constant, our error bound is tight up to a constant and matches with what is best possible for the unquantized observations. Next we comment on a number of remaining issues and possible extensions.

First, for the optimization, we leverage on a block relaxation algorithm. Although it can not guarantee the global optimality, our numerical experiments have suggested that the converged points often have nearly optimal objective values. In fact, following the proof of Theorem 1, the performance bound holds as long as the likelihood at $\hat{\Theta}$ is large enough, say,

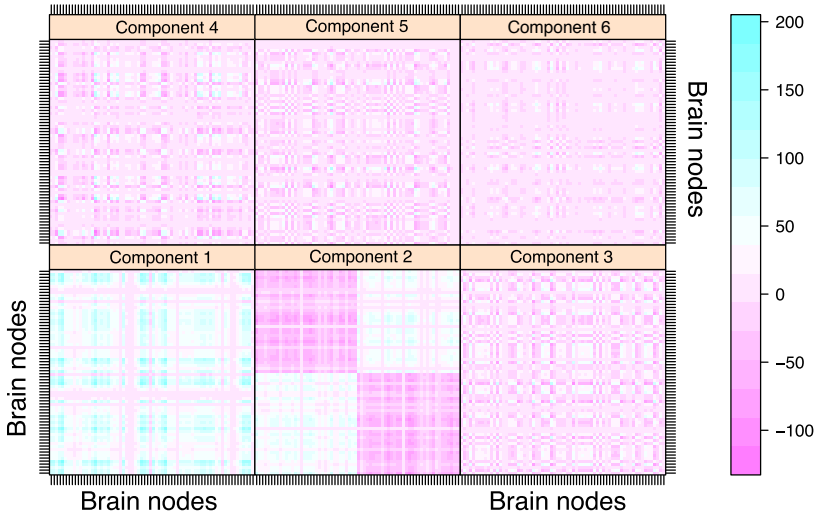


Figure 6: Analysis of the *HCP* data: heatmap for binary tensor components across brain regions. For each component, the connection matrix $\boldsymbol{A}_r = \lambda_r \boldsymbol{a}_r \otimes \boldsymbol{a}_r$ across the 68 brain nodes is plotted.

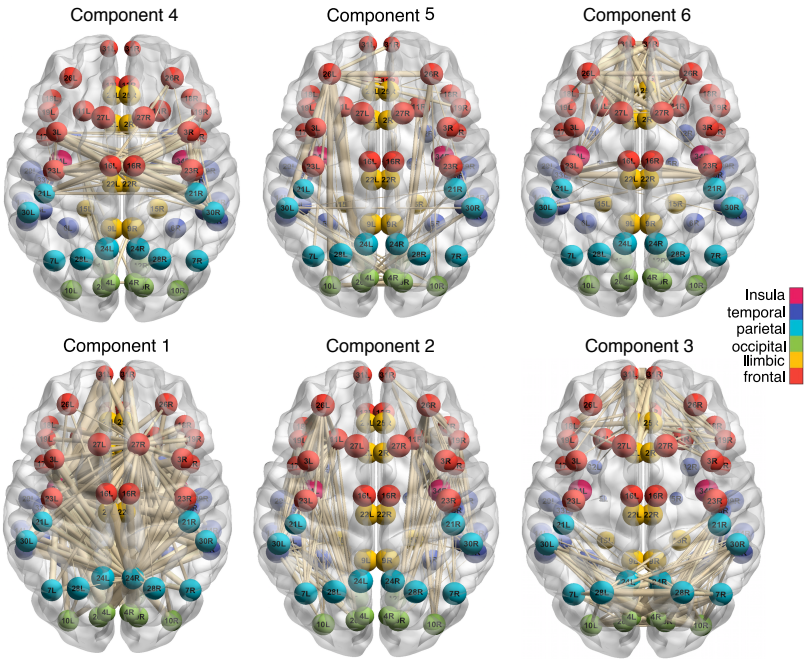


Figure 7: Analysis of the *HCP* data: edges with high loadings in the tensor decomposition components. The top 10% edges with positive $\boldsymbol{A}_r(i,j)$ are plotted. The width of the edge is proportional to the magnitude of $\boldsymbol{A}_r(i,j)$.

greater than or equal to $\mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}})$. When starting from random initializations, there could be multiple close-to-optimal choices of $\hat{\Theta}$, with negligible difference between their objective values. In that case, any of those choices performs equally well in estimating Θ_{true}. On the other hand, characterizing the global optimality for non-convex problems of this type has attracted a growing interest in the optimization community. Recent work has investigated non-convex optimization involving matrices and showed that there is no spurious local optimum for a certain type of objective (Ge et al., 2016). Non-convex optimization involving tensors are generally harder than matrices (Anandkumar et al., 2014), and its global optimality remains an open problem.

Second, for the theory, we assume the true rank R is known, whereas for the application, we propose to estimate the rank using BIC given the data. Actually, as long as the estimated rank is no smaller than the true rank, all our theoretical results still hold valid. On the other hand, it remains an open and challenging question to establish the convergence rate of the estimated rank (Zhou et al., 2013). We leave a full theoretical investigation of the rank selection consistency and the decomposition error bound under the estimated rank as future research.

Finally, although we have concentrated on the Bernoulli distribution in this article, we may consider extensions to other exponential-family distributions, so to account for count-valued tensors, multinomial-valued tensors, or tensors with mixed types of entries. Moreover, our proposed method can be thought of as a building block for more specialized tasks such as exploratory data analysis, tensor completion, compressed object representation, and network link prediction. Exploiting the benefits and properties of binary tensor decomposition in each specialized task warrants future research.

Appendix

A Technical lemmas

We begin with a set of technical lemmas that are useful for the proofs of the main theorems.

Lemma 1 (Tomioka and Suzuki (2014)). *Suppose that* $\mathcal{S} = [\![s_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ *is an order-K tensor whose entries are independent random variables that satisfy*

$$\mathbb{E}(s_{i_1,\ldots,i_K}) = 0, \quad \textit{and} \quad \mathbb{E}(e^{ts_{i_1,\ldots,i_K}}) \leq e^{t^2L^2/2}.$$

Then the spectral norm $\|\mathcal{S}\|_\sigma$ *satisfies that,*

$$\|\mathcal{S}\|_\sigma \leq \sqrt{8L^2 \log(12K) \sum_k d_k + \log(2/\delta)},$$

with probability at least $1 - \delta$.

Remark 1. The above lemma provides the bound on the spectral norm of random tensors. The result was firstly presented in Nguyen et al. (2015), and we adopt the version from Tomioka and Suzuki (2014).

Lemma 2. *Suppose that $\mathcal{S} = [\![s_{i_1,\ldots,i_K}]\!] \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is an order-K tensor whose entries are independent random variables that satisfy*

$$\mathbb{E}(s_{i_1,\ldots,i_K}) = 0, \quad \text{and} \quad |s_{i_1,\ldots,i_K}| \le L.$$

Then we have

$$\mathbb{P}\left(\|\mathcal{S}\|_\sigma \ge C_2 L\sqrt{\sum_k d_k}\right) \le \exp\left(-C_1 \log K \sum_k d_k\right)$$

where $C_1 > 0$ is an absolute constant, and $C_2 > 0$ is a constant that depends only on K.

Proof. Note that the random variable $L^{-1}s_{i_1,\ldots,i_K}$ is zero-mean and supported on $[-1,1]$. Therefore, $L^{-1}s_{i_1,\ldots,i_K}$ is sub-Gaussian with parameter $\frac{1-(-1)}{2} = 1$, i.e.

$$\mathbb{E}(L^{-1}s_{i_1,\ldots,i_K}) = 0, \quad \text{and} \quad \mathbb{E}(e^{tL^{-1}s_{i_1,\ldots,i_K}}) \le e^{t^2/2}.$$

It follows from Lemma 1 that, with probability at least $1-\delta$,

$$\|L^{-1}\mathcal{S}\|_\sigma \le \sqrt{(c_0 \log K + c_1)\sum_k d_k + \log(2/\delta)},$$

where $c_0, c_1 > 0$ are two absolute constants. Taking $\delta = \exp(-C_1 \log K \sum_k d_k)$ yields the final claim, where $C_2 = c_0 \log K + c_1 + 1 > 0$ is another constant. □

Lemma 3. *Let $\mathcal{A} \in \mathbb{R}^{d_1\times\cdots\times d_K}$ be an order-K tensor with rank$(\mathcal{A}) \le R$. Then*

$$\|\mathcal{A}\|_* \le R^{\frac{K-1}{2}}\|\mathcal{A}\|_F,$$

where $\|\cdot\|_*$ denotes the nuclear norm of the tensor.

Proof. Let rank$(\cdot)$ denote the regular matrix rank, and $\mathcal{A}_{(k)}$ denote the mode-k matricization of $\mathcal{A}$, $k \in [K]$. Define $\text{rank}_T(\mathcal{A}) = (R_1, \ldots, R_K)$ as the Tucker rank of $\mathcal{A}$, with $R_k = \text{rank}(\mathcal{A}_{(k)})$. The condition rank$(\mathcal{A}) \le R$ implies that $R_k \le R$ for all $k \in [K]$. Without loss of generality, assume $R_1 = \min_k R_k$. By Wang et al. (2017b, Corollary 4.11), and the invariance relationship between a tensor and its Tucker core (Jiang et al., 2017, Section 6), we have

$$\|\mathcal{A}\|_* \le \sqrt{\frac{\prod_{k=2}^K R_k}{\max_{k\ge 2} R_k}}\|\mathcal{A}_{(1)}\|_* \le R^{\frac{K-2}{2}}\|\mathcal{A}_{(1)}\|_*, \tag{13}$$

where $\mathcal{A}_{(1)}$ is a d_1-by-$\prod_{k\ge 2} d_k$ matrix with its rank bounded by R. Furthermore, the relationship between the matrix norms implies that $\|\mathcal{A}_{(1)}\|_* \le \sqrt{R}\|\mathcal{A}_{(1)}\|_F = \sqrt{R}\|\mathcal{A}\|_F$. Combining this fact with the inequality (13) yields the final claim. □

Lemma 4. *Let $\mathcal{A}, \mathcal{B}$ be two order-K tensors of the same dimension. Then*

$$|\langle \mathcal{A}, \mathcal{B}\rangle| \le \|\mathcal{A}\|_\sigma \|\mathcal{B}\|_*.$$

Proof. By Friedland and Lim (2018, Proposition 3.1), there exists a nuclear norm decomposition of $\mathcal{B}$, such that

$$\mathcal{B} = \sum_r \lambda_r \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)}, \quad \boldsymbol{a}_r^{(k)} \in \mathbf{S}^{d_k - 1}(\mathbb{R}), \quad \text{for all } k \in [K],$$

and $\|\mathcal{B}\|_* = \sum_r |\lambda_r|$. Henceforth we have

$$\begin{aligned} |\langle \mathcal{A}, \mathcal{B} \rangle| &= |\langle \mathcal{A}, \sum_r \lambda_r \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)} \rangle| \leq \sum_r |\lambda_r| |\langle \mathcal{A}, \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(K)} \rangle| \\ &\leq \sum_r |\lambda_r| \|\mathcal{A}\|_\sigma = \|\mathcal{A}\|_\sigma \|\mathcal{B}\|_*, \end{aligned}$$

which completes the proof. □

Lemma 5. *Let X, Y be two Bernoulli random variables with means p and q, $0 < p, q < 1$, respectively. Then the Kullback-Leibler (KL) divergence satisfies that*

$$KL(X, Y) \leq \frac{(p-q)^2}{q(1-q)},$$

where $KL(X, Y) = -\sum_{x=\{0,1\}} P_X(x) \log \left\{ \frac{P_Y(x)}{P_X(x)} \right\}$.

Proof. It is straightforward to verify that,

$$KL(X, Y) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \ \leq \ p \frac{p-q}{q} + (1-p) \frac{q-p}{1-q} = \frac{(p-q)^2}{q(1-q)},$$

where the inequality is due to the fact that $\log x \leq x - 1$ for $x > 0$. □

Lemma 6. *Let $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$ be a binary tensor. Let $\mathbb{P}_\Theta$ denote the distribution of $\mathcal{Y}|\Theta$ based on the Bernoulli model (2) with the link function f and the parameter tensor Θ. Let $\mathbb{P}_{\mathbf{0}}$ denote the distribution of $\mathcal{Y}|\mathbf{0}$ induced by the zero parameter tensor. Then*

$$KL(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq 4 \dot{f}^2(0) \|\Theta\|_F^2.$$

Proof. We have that

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) &= \sum_{i_1, \ldots, i_K} \mathrm{KL}(\mathcal{Y}_{i_1, \ldots, i_K} | \theta_{i_1, \ldots, i_K}, \mathcal{Y}_{i_1, \ldots, i_K} | 0) \leq \sum_{i_1, \ldots, i_K} \frac{(f(\theta_{i_1, \ldots, i_K}) - f(0))^2}{f(0)(1 - f(0))} \\ &= \sum_{i_1, \ldots, i_K} \frac{\dot{f}^2(\eta_{i_1, \ldots, i_K} \theta_{i_1, \ldots, i_K}) \, (\theta_{i_1, \ldots, i_K} - 0)^2}{f(0)(1 - f(0))} \leq \sum_{i_1, \ldots, i_K} 4 \dot{f}^2(0) \theta_{i_1, \ldots, i_K}^2 \\ &= 4 \dot{f}^2(0) \|\Theta\|_F^2, \end{aligned}$$

where the first inequality comes from Lemma 5, the next equality comes from the first-order Taylor expansion with $\eta_{i_1, \ldots, i_K} \in [0, 1]$, and the last inequality uses the fact that $f(0) = 1/2$ and f' peaks at zero for an unimodal and symmetric density function. □

Lemma 7 (Varshamov-Gilbert bound). *Let $\Omega = \{(w_1, \ldots, w_m) : w_i \in \{0, 1\}\}$. Suppose $m > 8$. Then there exists a subset $\{w^{(0)}, \ldots, w^{(M)}\}$ of Ω such that $w^{(0)} = (0, \ldots, 0)$ and*

$$\|w^{(j)} - w^{(k)}\|_0 \geq \frac{m}{8}, \quad \text{for } 0 \leq j < k \leq M,$$

where $\|\cdot\|_0$ denotes the Hamming distance, and $M \geq 2^{m/8}$.

Lemma 8. *Assume the same setup as in Theorem 2. Without loss of generality, suppose $d_1 = d_{\max}$, and define $d_{total} = \prod_{k \geq 1} d_k$. For any constant $0 \leq \gamma \leq 1$, there exist a finite set of tensors $\mathcal{X} = \{\Theta_i : i = 1, \ldots\} \subset \mathcal{D}(R, \alpha)$ satisfying the following four properties:*

(i) Card$(\mathcal{X}) \geq 2^{Rd_1/8} + 1$, where Card denotes the cardinality;

(ii) $\mathcal{X}$ contains the zero tensor $\mathbf{0} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$;

(iii) $\|\Theta\|_\infty \leq \gamma \min\left\{\alpha, \sigma\sqrt{\frac{Rd_1}{d_{total}}}\right\}$ for any element $\Theta \in \mathcal{X}$;

(iv) $\|\Theta_i - \Theta_j\|_F \geq \frac{\gamma}{4} \min\left\{\alpha\sqrt{d_{total}}, \sigma\sqrt{Rd_1}\right\}$ for any two distinct elements $\Theta_i, \Theta_j \in \mathcal{X}$.

Proof. Given a constant $0 \leq \gamma \leq 1$, we define a set of matrices:

$$\mathcal{C} = \left\{ \boldsymbol{M} = (m_{ij}) \in \mathbb{R}^{d_1 \times R} : a_{ij} \in \left\{0, \gamma \min\left\{\alpha, \sigma\sqrt{\frac{Rd_1}{d_{\text{total}}}}\right\}\right\},\ \forall (i, j) \in [d_1] \times [R] \right\}.$$

We then consider the associated set of block tensors:

$$\mathcal{B} = \mathcal{B}(\mathcal{C}) = \big\{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \Theta(\cdot,\ \cdot,\ i_3, \ldots, i_K) = (\boldsymbol{M}|\cdots|\boldsymbol{M}|\boldsymbol{O}) \in \mathbb{R}^{d_1 \times d_2},$$
$$\text{for all } (i_3, \ldots, i_K) \in [d_3] \times \cdots \times [d_K],\ \text{where } \boldsymbol{M} \in \mathcal{C}\big\},$$

where $\Theta(\cdot,\ \cdot,\ i_3, \ldots, i_K)$ denotes the matrix obtained from Θ by varying the indices in the first two modes and fixing the indices in the other modes, $\boldsymbol{O}$ denotes the $d_1 \times (d_2 - R\lfloor d_2/R \rfloor)$ zero matrix, and $\lfloor d_2/R \rfloor$ is the integer part of d_2/R. In other words, the tensor Θ consists of blocks of the type $\Theta(\cdot,\ \cdot,\ i_3, \ldots, i_K)$ for any $(i_3, \ldots, i_K) \in [d_3] \times \cdots \times [d_K]$, stacked on top of each other, and furthermore, the block $\Theta(\cdot,\ \cdot,\ i_3, \ldots, i_K) \in \mathbb{R}^{d_1 \times d_2}$ is filled out by copying the matrix $\boldsymbol{M} \in \mathbb{R}^{d_1 \times R}$ as many times as would fit.

By construction, any element of $\mathcal{B}$, as well as the difference of any two elements of $\mathcal{B}$, has tensor rank at most R, and the entries of any tensor in $\mathcal{B}$ take values in $[0, \alpha]$. Thus, $\mathcal{B} \subset \mathcal{D}(R, \alpha)$. By Lemma 7, there exists a subset $\mathcal{X} \subset \mathcal{B}$ with cardinality Card$(\mathcal{X}) \geq 2^{d_1 R/8} + 1$ containing the zero $d_1 \times \cdots d_K$ tensor, such that, for any two distinct elements Θ_i and Θ_j in $\mathcal{X}$,

$$\|\Theta_i - \Theta_j\|_F^2 \geq \frac{d_1 R}{8}\gamma^2 \min\left\{\alpha, \frac{\sigma^2 R d_1}{d_{\text{total}}}\right\} \lfloor \frac{d_2}{R} \rfloor \prod_{k \geq 3} d_k \geq \frac{\gamma^2 \min\left\{\alpha^2 d_{\text{total}}, \sigma^2 R d_1\right\}}{16}.$$

In addition, each entry of $\Theta \in \mathcal{X}$ is bounded by $\gamma \min\left\{\alpha, \sigma\sqrt{\frac{Rd_1}{d_{\text{total}}}}\right\}$. Therefore the Properties (i) to (iv) are satisfied. □

B Proofs

B.1 Proof of Theorem 1

Proof. It follows from the expression of $\mathcal{L}_{\mathcal{Y}}(\Theta)$ that

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\partial \theta_{i_1,\ldots,i_K}} &= \frac{\dot{f}(\theta_{i_1,\ldots,i_K})}{f(\theta_{i_1,\ldots,i_K})}\mathbb{1}_{\{y_{i_1,\ldots,i_K}=1\}} - \frac{\dot{f}(\theta_{i_1,\ldots,i_K})}{1-f(\theta_{i_1,\ldots,i_K})}\mathbb{1}_{\{y_{i_1,\ldots,i_K}=-1\}},\\
\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial \theta^2_{i_1,\ldots,i_K}} &= -\left[\frac{\dot{f}^2(\theta_{i_1,\ldots,i_K})}{f^2(\theta_{i_1,\ldots,i_K})} - \frac{\ddot{f}(\theta_{i_1,\ldots,i_K})}{f(\theta_{i_1,\ldots,i_K})}\right]\mathbb{1}_{\{y_{i_1,\ldots,i_K}=1\}}\\
&\quad - \left[\frac{\ddot{f}(\theta_{i_1,\ldots,i_K})}{1-f(\theta_{i_1,\ldots,i_K})} + \frac{\dot{f}^2(\theta_{i_1,\ldots,i_K})}{\{1-f(\theta_{i_1,\ldots,i_K})\}^2}\right]\mathbb{1}_{\{y_{i_1,\ldots,i_K}=-1\}},\\
\frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial \theta_{i_1,\ldots,i_K}\theta_{i'_1,\ldots,i'_K}} &= 0, \quad \text{if} \quad (i_1,\ldots,i_K) \neq (i'_1,\ldots,i'_K).
\end{aligned}$$

Define

$$\mathcal{S}_{\mathcal{Y}}(\Theta_{\text{true}}) = \left[\!\!\left[\frac{\partial \mathcal{L}_{\mathcal{Y}}}{\theta_{i_1,\ldots,i_K}}\right]\!\!\right]\Big|_{\Theta=\Theta_{\text{true}}}, \quad \text{and} \quad \mathcal{H}_{\mathcal{Y}}(\Theta_{\text{true}}) = \left[\!\!\left[\frac{\partial \mathcal{L}^2_{\mathcal{Y}}}{\partial\theta_{i_1,\ldots,i_K}\theta_{i'_1,\ldots,i'_K}}\right]\!\!\right]\Big|_{\Theta=\Theta_{\text{true}}},$$

where the former is the collection of the score functions evaluated at Θ_{true}, and the latter is the collection of the Hession functions evaluated at Θ_{true}. We organize the entries in $\mathcal{S}_{\mathcal{Y}}(\Theta_{\text{true}})$ and treat $\mathcal{S}_{\mathcal{Y}}(\Theta_{\text{true}})$ as an order-K dimension-$(d_1,\ldots,d_K)$ tensor. Similarly, we organize the entries in $\mathcal{H}_{\mathcal{Y}}(\Theta_{\text{true}})$ and treat $\mathcal{H}_{\mathcal{Y}}(\Theta_{\text{true}})$ as a $\prod_k d_k$-by-$\prod_k d_k$ matrix. By the second-order Taylor's theorem, we expand $\mathcal{L}_{\mathcal{Y}}(\Theta)$ around Θ_{true} and obtain

$$\mathcal{L}_{\mathcal{Y}}(\Theta) = \mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}}) + \langle S_{\mathcal{Y}}(\Theta_{\text{true}}), \Theta - \Theta_{\text{true}}\rangle + \frac{1}{2}\operatorname{vec}(\Theta-\Theta_{\text{true}})^T \mathcal{H}_{\mathcal{Y}}(\check{\Theta})\operatorname{vec}(\Theta-\Theta_{\text{true}}), \quad (14)$$

where $\check{\Theta} = \gamma\Theta_{\text{true}} + (1-\gamma)\Theta$ for some $\gamma \in [0,1]$, and $\mathcal{H}_{\mathcal{Y}}(\check{\Theta})$ denotes the $\prod_k d_k$-by-$\prod_k d_k$ Hession matrix evaluated at $\check{\Theta}$.

We first bound the linear term in (14). Note that, by Lemma 4,

$$|\langle S_{\mathcal{Y}}(\Theta_{\text{true}}), \Theta - \Theta_{\text{true}}\rangle| \leq \|S_{\mathcal{Y}}(\Theta_{\text{true}})\|_\sigma \|\Theta - \Theta_{\text{true}}\|_*. \quad (15)$$

Define

$$s_{i_1,\ldots,i_K} = \frac{\partial \mathcal{L}_{\mathcal{Y}}}{\theta_{i_1,\ldots,i_K}}|_{\Theta=\Theta_{\text{true}}} \quad \text{for all} \quad (i_1,\ldots,i_K) \in [d_1]\times\cdots\times[d_K].$$

It follows from model (2) and the expression for L_α that $S_{\mathcal{Y}}(\Theta_{\text{true}}) = [\![s_{i_1,\ldots,i_K}]\!]$ is a random tensor whose entries are independently distributed and satisfy

$$\mathbb{E}(s_{i_1,\ldots,i_K}) = 0, \quad |s_{i_1,\ldots,i_K}| \leq L_\alpha, \quad \text{for all } (i_1,\ldots,i_K) \in [d_1]\times\cdots\times[d_K]. \quad (16)$$

By lemma 2, with probability at least $1-\exp(-C_1 \log K \sum_k d_k)$, we have

$$\|S_{\mathcal{Y}}(\Theta_{\text{true}})\|_\sigma \leq C_2 L_\alpha \log K \sqrt{\sum_k d_k}, \quad (17)$$

where C_1, C_2 are two positive constants. Furthermore, note that $\text{rank}(\Theta) \le R$, $\text{rank}(\Theta_{\text{true}}) \le R$, so $\text{rank}(\Theta - \Theta_{\text{true}}) \le 2R$. By lemma 3, $\|\Theta - \Theta_{\text{true}}\|_* \le (2R)^{\frac{K-1}{2}} \|\Theta - \Theta_{\text{true}}\|_F$. Combining (15), (16) and (17), we have that, with probability at least $1 - \exp(-C_1 \log K \sum_k d_k)$,

$$|\langle S_{\mathcal{Y}}(\Theta_{\text{true}}), \Theta - \Theta_{\text{true}} \rangle| \le C_2 L_\alpha \sqrt{R^{K-1} \sum_k d_k} \|\Theta - \Theta_{\text{true}}\|_F, \tag{18}$$

where the constant C_2 absorbs all factors that depend only on K.

We next bound the quadratic term in (14). Let $W = \text{vec}(\Theta - \Theta_{\text{true}})$. Note that

$$\begin{aligned} \text{vec}(\Theta - \Theta_{\text{true}})^T H_{\mathcal{Y}}(\check{\Theta}) \, \text{vec}(\Theta - \Theta_{\text{true}}) &= \sum_{i_1,\ldots,i_K} \left(\frac{\partial \mathcal{L}_{\mathcal{Y}}^2}{\partial \theta_{i_1,\ldots,i_K}^2} \Big|_{\Theta = \check{\Theta}} \right) (\Theta_{i_1,\ldots,i_K} - \Theta_{\text{true},i_1,\ldots,i_K})^2 \\ &\le -\gamma_\alpha \sum_{i_1,\ldots,i_K} (\Theta_{i_1,\ldots,i_K} - \Theta_{\text{true},i_1,\ldots,i_K})^2 \\ &= -\gamma_\alpha \|\Theta - \Theta_{\text{true}}\|_F^2, \end{aligned} \tag{19}$$

where the second line comes from the fact that $\|\check{\Theta}\|_\infty \le \alpha$ and the definition of γ_α.

Combining (14), (18) and (19), we have that, for all $\Theta \in \mathcal{D}$, with probability at least $1 - \exp(-C_1 \log K \sum_k d_k)$,

$$\mathcal{L}_{\mathcal{Y}}(\Theta) \le \mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}}) + C_2 L_\alpha \left(R^{K-1} \sum_k d_k \right)^{1/2} \|\Theta - \Theta_{\text{true}}\|_F - \frac{\gamma_\alpha}{2} \|\Theta - \Theta_{\text{true}}\|_F^2,$$

In particular, this also holds for $\hat{\Theta} \in \mathcal{D}$, in that

$$\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}) \le \mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}}) + C_2 L_\alpha \left(R^{K-1} \sum_k d_k \right)^{1/2} \|\hat{\Theta} - \Theta_{\text{true}}\|_F - \frac{\gamma_\alpha}{2} \|\hat{\Theta} - \Theta_{\text{true}}\|_F^2.$$

Since $\hat{\Theta} = \arg\max_{\Theta \in \mathcal{D}} \mathcal{L}_{\mathcal{Y}}$, $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}) - \mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}}) \ge 0$, which gives

$$C_2 L_\alpha \left(R^{K-1} \sum_k d_k \right)^{1/2} \|\hat{\Theta} - \Theta_{\text{true}}\|_F - \frac{\gamma_\alpha}{2} \|\hat{\Theta} - \Theta_{\text{true}}\|_F^2 \ge 0.$$

Henceforth,

$$\frac{1}{\sqrt{\prod_k d_k}} \|\hat{\Theta} - \Theta_{\text{true}}\|_F \le \frac{2 C_2 L_\alpha \sqrt{R^{K-1} \sum_k d_k}}{\gamma_\alpha \sqrt{\prod_k d_k}} = 2 C_2 \frac{L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}}.$$

This completes the proof. □

Remark 2. Based on the proof of Theorem 1, we can relax the MLE assumption on the estimator $\hat{\Theta}$. The same convergence rate holds in the level set $\{\Theta \in \mathcal{D} : \mathcal{L}_{\mathcal{Y}}(\Theta) \ge \mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}})\}$.

B.2 Proof of Theorem 2

Proof. Without loss of generality, we assume $d_1 = d_{\max}$, and denote by $d_{\text{total}} = \prod_{k\geq 1} d_k$. Let $\gamma \in [0,1]$ be a constant to be specified later. Our strategy is to construct a finite set of tensors $\mathcal{X} = \{\Theta_i : i = 1, \cdots\} \subset \mathcal{D}(R,\alpha)$ satisfying the properties of (i)-(iv) in Lemma 8. By Lemma 8, such a subset of tensors exist. For any tensor $\Theta \in \mathcal{X}$, let $\mathbb{P}_\Theta$ denote the distribution of $\mathcal{Y}|\Theta$, where $\mathcal{Y}$ is the observed binary tensor. In particular, $\mathbb{P}_{\mathbf{0}}$ is the distribution of $\mathcal{Y}$ induced by the zero parameter tensor $\mathbf{0}$, i.e., the distribution of $\mathcal{Y}$ conditional on the coefficient tensor $\Theta = \mathbf{0}$. Then conditioning on $\Theta \in \mathcal{X}$, the entries of $\mathcal{Y}$ are independent Bernoulli random variables. In addition, we note that (c.f. Lemma 5),

$$\begin{aligned} \text{for the logistic link:} \quad & \mathrm{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \frac{4}{\sigma^2}\|\Theta\|_F^2, \\ \text{for the probit link:} \quad & \mathrm{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \frac{2}{\pi\sigma^2}\|\Theta\|_F^2, \\ \text{for the Laplacian link:} \quad & \mathrm{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \frac{1}{\sigma^2}\|\Theta\|_F^2, \end{aligned} \tag{20}$$

where σ is the scale parameter. Therefore, we have that the KL divergence between $\mathbb{P}_\Theta$ and $\mathbb{P}_{\mathbf{0}}$ satisfies

$$\mathrm{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \frac{2}{\pi\sigma^2}\|\Theta\|_F^2 \leq \frac{1}{8\pi} R d_1 \gamma^2, \tag{21}$$

where the first inequality comes from (20), and the second inequality comes from property (iv) of $\mathcal{X}$. From (21) and the property (i), we deduce that the condition

$$\frac{1}{\mathrm{Card}(\mathcal{X}) - 1} \sum_{\Theta \in \mathcal{X}} \mathrm{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \varepsilon \log\{\mathrm{Card}(\mathcal{X}) - 1\} \tag{22}$$

is satisfied for any $\varepsilon \geq 0$ if $\gamma \in [0,1]$ is chosen to be sufficiently small, depending on ε, e.g., $\gamma \leq \sqrt{3\varepsilon}$. By applying Tsybakov (2009, Theorem 2.5) to (22), and in view of the property (iv), we obtain that

$$\inf_{\hat{\Theta}} \sup_{\Theta_{\text{true}} \in \mathcal{X}} \mathbb{P}\left(\|\hat{\Theta} - \Theta_{\text{true}}\|_F \geq \frac{\gamma}{8} \min\left\{ \alpha\sqrt{d_{\text{total}}}, \sigma\sqrt{R d_1} \right\} \right) \geq \frac{1}{2}\left(1 - 2\varepsilon - \sqrt{\frac{16\varepsilon}{R d_1}} \right). \tag{23}$$

Note that $\|\hat{\Theta} - \Theta_{\text{true}}\|_F = \sqrt{d_{\text{total}}}\mathrm{Loss}(\hat{\Theta}, \Theta_{\text{true}})$ and $\mathcal{X} \subset \mathcal{D}(R,\alpha)$. Therefore, we conclude from (23) that

$$\inf_{\hat{\Theta}} \sup_{\Theta_{\text{true}} \in \mathcal{D}(R,\alpha)} \mathbb{P}\left(\mathrm{Loss}(\hat{\Theta}, \Theta_{\text{true}}) \geq \frac{1}{16} \min\left\{ \alpha, \sigma\sqrt{\frac{R d_{\max}}{d_{\text{total}}}} \right\} \right) \geq \frac{1}{2}\left(\frac{3}{4} - \sqrt{\frac{2}{R d_{\max}}} \right) \geq \frac{1}{8},$$

by taking $\varepsilon = 1/8$ and $\gamma = 1/2$. This completes the proof. □

B.3 Proof of Theorem 3

Proof. The argument is similar as that in the proof of Theorem 2. Specifically, we construct a set of tensors $\mathcal{X} \subset \mathcal{D}(R,\alpha)$ such that, for any $\Theta \in \mathcal{X}$, Θ satisfies the properties (i) to (iv) of Lemma 8. Given a continuous-valued tensor $\mathcal{Y}$, let $\mathbb{P}_\Theta$ denote the distribution of $\mathcal{Y}|\Theta$

according to the Gaussian model; that is, $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!]|\Theta \sim_{\text{i.i.d.}} N(0,\sigma^2)$. Note that, for the Gaussian distribution,

$$KL(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) = \frac{\|\Theta\|_F}{2\sigma^2} \leq \frac{1}{16} R d_1 \gamma^2.$$

So the condition

$$\frac{1}{\text{Card}(\mathcal{X}) - 1} \sum_{\Theta \in \mathcal{X}} \text{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \varepsilon \log\left(\text{Card}(\mathcal{X}) - 1\right) \tag{24}$$

is satisfied for any $\varepsilon \geq 0$ if $\gamma \in [0,1]$ is chosen to be sufficiently small, depending on ε, e.g., $\gamma \leq \sqrt{3\varepsilon}$. In view of the property (iv) and (24), the conclusion now follows from the application of Tsybakov (2009, Theorem 2.5). This completes the proof. □

B.4 Proof of Proposition 1

Proof. The proof of the global convergence is similar to that of Zhou et al. (2013, Proposition 1). We present the main ideas here for completeness. By Assumption (A2), the block update is well-defined and differentiable. The isolation of stationary points ensures that there are only finite number of stationary points. It sufficies to show that the every sub-sequence of $\boldsymbol{A}^{(t)}$ convergences to a same limiting point; i.e.,

$$\limsup_{t\to\infty} \boldsymbol{A}^{(t)} = \liminf_{t\to\infty} \boldsymbol{A}^{(t)}. \tag{25}$$

Let $\boldsymbol{A}^{(t_n)}$ be a sub-sequence with limiting point $\boldsymbol{A}^*$. As the algorithm monotonically increases the objective value, the limiting point $\boldsymbol{A}^*$ is a stationary point of $\mathcal{L}$. Now take the set of all limiting points, which is contained in the set $\{\boldsymbol{A}\colon \mathcal{L}(\boldsymbol{A}) \geq \mathcal{L}(\boldsymbol{A}^{(0)})\}$, and is thus compact due to (A1). The compactness implies that the set of limiting points is also connected (Lange, 2012, Propositions 8.2.1 and 15.4.2). Therefore the set of limiting points is a connected subset of the finite stationery points, and thus becomes a single point. In other words, (25) holds and $\boldsymbol{A}^{(t)}$ convergences to a stationary point of $\mathcal{L}$.

The local convergence follows from Uschmajew (2012, Theorem 3.3) and Zhou et al. (2013, Proposition 1). Here we elaborate on the contraction parameter $\rho \in (0,1)$ in our context. Let $\boldsymbol{H}$ denote the Hession matrix of the log-likelihood $\mathcal{L}(\boldsymbol{A})$ at the local maximum $\boldsymbol{A}^*$. We partition the Hession into $\boldsymbol{H} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{L}^T$, where $\boldsymbol{L}$ is the strictly block lower triangular part and $\boldsymbol{D}$ is the block diagonal part. By Assumption (A2), each sub-block of the Hession is negative definite, so the diagonal entries of $\boldsymbol{D}$ are strictly negative. This ensures that the block lower triangular matrix $\boldsymbol{L} + \boldsymbol{D}$ is invertible. The differential of the algorithm map $\mathcal{M} : \boldsymbol{A}^{(t)} \mapsto \boldsymbol{A}^{(t+1)}$ can be shown as $\mathcal{M}' = -(\boldsymbol{L} + \boldsymbol{D})^{-1}\boldsymbol{L}$ (Bezdek and Hathaway, 2003, Lemma 2). Therefore $\rho = \max_i \left|\lambda_i\left\{(\boldsymbol{L} + \boldsymbol{D})^{-1}\boldsymbol{L}\right\}\right| \in (0,1)$. By the contraction principle,

$$\|\boldsymbol{A}^{(t)} - \boldsymbol{A}^*\|_F \leq \rho^t \|\boldsymbol{A}^{(0)} - \boldsymbol{A}^*\|_F,$$

for $\boldsymbol{A}^{(0)}$ sufficiently close to $\boldsymbol{A}^*$. Because $\Theta = \Theta(\boldsymbol{A})$ is local Lipschitz at $\boldsymbol{A}^*$ with constants $c_1, c_2 > 0$,

$$c_1\|\boldsymbol{A}^{(t)} - \boldsymbol{A}^*\|_F \leq \|\Theta(\boldsymbol{A}^{(t)}) - \Theta(\boldsymbol{A}^*)\|_F \leq c_2\|\boldsymbol{A}^{(t)} - \boldsymbol{A}^*\|_F,$$

for any sufficiently large $t \in \mathbb{N}_+$. Therefore

$$\|\Theta(\boldsymbol{A}^{(t)}) - \Theta^*\|_F \leq \rho^t C \|\Theta(\boldsymbol{A}^{(0)}) - \Theta^*\|_F,$$

where $C > 0$ is a constant. □

B.5 Proof of Theorem 4

Proof. Based on Remark 2 after Theorem 1, we have

$$\text{Loss}(\Theta^*, \Theta_{\text{true}}) \leq \frac{C_2 L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}}.$$

Meanwhile, Proposition 1 implies that

$$\text{Loss}(\Theta^{(t)}, \Theta^*) \leq C_1 \rho^t \text{Loss}(\Theta^{(0)}, \Theta^*).$$

Combining the above two results yields

$$\begin{aligned} \text{Loss}(\Theta^{(t)}, \Theta_{\text{true}}) &\leq \text{Loss}(\Theta^{(t)}, \Theta^*) + \text{Loss}(\Theta_{\text{true}}, \Theta^*) \\ &\leq C_1 \rho^t \text{Loss}(\Theta^{(0)}, \Theta^*) + \frac{C_2 L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}}. \end{aligned}$$

This completes the proof. □

Acknowledgements

Wang's research was partially supported by the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. Li's research was partially supported by NSF grant DMS-1613137 and NIH grants R01AG034570 and R01AG061303. The authors thank the Associate Editor and three referees for their constructive comments.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Evrim Acar, Daniel M Dunlavy, Tamara G Kolda, and Morten Mørup. Scalable tensor factorizations with missing data. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 701–712. SIAM, 2010.

Adelin Albert and John A Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

James C Bezdek and Richard J Hathaway. Convergence of alternating optimization. *Neural, Parallel & Scientific Computations*, 11(4):351–368, 2003.

Sonia A Bhaskar and Adel Javanmard. 1-bit matrix completion under exact low-rank constraint. In *Information Sciences and Systems (CISS), 2015 49th Annual Conference on*, pages 1–6. IEEE, 2015.

Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. Uniqueness of tensor decompositions with applications to polynomial identifiability. In *Conference on Learning Theory*, pages 742–778, 2014.

Xuan Bi, Annie Qu, and Xiaotong Shen. Multilayer tensor factorization with applications to recommender systems. *Ann. Statist.*, 46(6B):3308–3333, 12 2018. doi: 10.1214/17-AOS1659.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

Han Chen, Garvesh Raskutti, and Ming Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *arXiv preprint arXiv:1611.10349*, 2016.

Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624, 2002.

Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

Jan De Leeuw. Principal component analysis of binary data by iterated singular value decomposition. *Computational statistics & data analysis*, 50(1):21–39, 2006.

Woodrow W Denham and Douglas R White. Multiple measures of Alyawarra kinship. *Field Methods*, 17(1):70–101, 2005.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

Dóra Erdos and Pauli Miettinen. Discovering facts with boolean tensor tucker decomposition. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1569–1572. ACM, 2013a.

Dóra Erdos and Pauli Miettinen. Walk'n'merge: a scalable algorithm for boolean tensor factorization. In *2013 IEEE 13th International Conference on Data Mining*, pages 1037–1042. IEEE, 2013b.

Shmuel Friedland and Lek-Heng Lim. Nuclear norm of higher-order tensors. *Mathematics of Computation*, 87(311):1255–1281, 2018.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Navid Ghadermarzy, Yaniv Plan, and Ozgur Yilmaz. Learning tensors from partial binary measurements. *arXiv preprint arXiv:1804.00108*, 2018.

Ilan I Goldberg, Michal Harel, and Rafael Malach. When the brain loses its self: prefrontal inactivation during sensorimotor processing. *Neuron*, 50(2):329–339, 2006.

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

David Hong, Tamara G Kolda, and Jed A Duersch. Generalized canonical polyadic tensor decomposition. *arXiv preprint arXiv:1808.07452*, 2018.

Victoria Hore, Ana Viñuela, Alfonso Buil, Julian Knight, Mark I McCarthy, Kerrin Small, and Jonathan Marchini. Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094, 2016.

Qinghao Hu, Gang Li, Peisong Wang, Yifan Zhang, and Jian Cheng. Training binary weight networks via semi-binary decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018.

Bo Jiang, Fan Yang, and Shuzhong Zhang. Tensor and its Tucker core: the invariance relationships. *Numerical Linear Algebra with Applications*, 24(3):e2086, 2017.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

Kenneth Lange. *Numerical Analysis for Statisticians*. Springer Publishing Company, Incorporated, 2nd edition, 2012. ISBN 146142612X, 9781461426127.

Seokho Lee, Jianhua Z Huang, and Jianhua Hu. Sparse logistic principal components analysis for binary data. *The annals of applied statistics*, 4(3):1579, 2010.

Jakub Mažgut, Peter Tiňo, Mikael Bodén, and Hong Yan. Dimensionality reduction and topographic mapping of binary tensors. *Pattern Analysis and Applications*, 17(3):497–515, 2014.

Pauli Miettinen. Boolean tensor factorizations. In *2011 IEEE 11th International Conference on Data Mining*, pages 447–456. IEEE, 2011.

Cun Mu, Bo Huang, John Wright, and Donald Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, pages 73–81, 2014.

Nam H Nguyen, Petros Drineas, and Trac D Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *Information and Inference: A Journal of the IMA*, 4(3):195–229, 2015.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*, volume 11, pages 809–816, 2011.

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

Piyush Rai, Yingjian Wang, Shengbo Guo, Gary Chen, David Dunson, and Lawrence Carin. Scalable bayesian low-rank decomposition of incomplete multiway tensors. In *International Conference on Machine Learning*, pages 1800–1808, 2014.

Piyush Rai, Changwei Hu, Matthew Harding, and Lawrence Carin. Scalable probabilistic tensor factorization for binary and count data. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 3770–3776, 2015.

Jarod L Roland, Abraham Z Snyder, Carl D Hacker, Anish Mitra, Joshua S Shimony, David D Limbrick, Marcus E Raichle, Matthew D Smyth, and Eric C Leuthardt. On the role of the corpus callosum in interhemispheric functional connectivity in humans. *Proceedings of the National Academy of Sciences*, 114(50):13278–13283, 2017.

Tammo Rukat, Chris Holmes, and Christopher Yau. Probabilistic boolean tensor decomposition. In *International conference on machine learning*, pages 4410–4419, 2018.

Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 899–916, 2017. doi: 10.1111/rssb.12190.

Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton. Tensor analyzers. In *International Conference on Machine Learning*, pages 163–171, 2013.

Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.

Alexandre B Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by Vladimir Zaiats, 2009.

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

André Uschmajew. Local convergence of the alternating least squares algorithm for canonical tensor approximation. *SIAM Journal on Matrix Analysis and Applications*, 33(2): 639–652, 2012.

Lu Wang, Zhengwu Zhang, and David Dunson. Common and individual structure of multiple networks. *arXiv preprint arXiv:1707.06360*, 2017a.

Miaoyan Wang and Yun Song. Tensor decompositions via two-mode higher-order SVD (HOSVD). In *Artificial Intelligence and Statistics*, pages 614–622, 2017.

Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear algebra and its applications*, 520:44–66, 2017b.

Miaoyan Wang, Jonathan Fischer, and Yun S Song. Three-way clustering of multi-tissue multi-individual gene expression data using constrained tensor decomposition. *bioRxiv 229245*, 2017c.

Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

Anru Zhang and Dong Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 2018.

Zhengwu Zhang, Maxime Descoteaux, Jingwen Zhang, Gabriel Girard, Maxime Chamberland, David Dunson, Anuj Srivastava, and Hongtu Zhu. Mapping population-based structural connectomes. *NeuroImage*, 172:130–145, 2018.

Shandian Zhe, Kai Zhang, Pengyuan Wang, Kuang-chih Lee, Zenglin Xu, Yuan Qi, and Zoubin Ghahramani. Distributed flexible nonlinear tensor factorization. In *Advances in Neural Information Processing Systems*, pages 928–936, 2016.

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.