

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

**Miaoyan Wang**

*Department of Statistics  
University of Wisconsin-Madison  
Madison, WI 53706, USA*

MIAOYAN.WANG@WISC.EDU

**Lexin Li**

*Department of Biostatistics and Epidemiology  
University of California-Berkeley  
Berkeley, CA 94720-1776, USA*

LEXINLI@BERKELEY.EDU

**Editor:**

## 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, Generalized linear model, Constrained maximum likelihood estimation, High-dimensionality.

## 1. Introduction

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 for such an increase 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. We propose to consider the following low-rank Bernoulli model:

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

where  $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!] \in \{0, 1\}^{d_1 \times \dots \times d_K}$  is a binary tensor, whose entries are either 1 or 0, encoding the presence or absence of the event indexed by the  $K$ -tuple  $(i_1, \dots, i_K)$ . The entries of  $\mathcal{Y}$  are assumed to be realizations of independent Bernoulli random variables with success probability  $f(\theta_{i_1, \dots, i_K})$ . Here  $f$  is a suitable chosen function that maps  $\mathbb{R}$  to  $[0, 1]$ . The parameter tensor,  $\Theta = [\![\theta_{i_1, \dots, i_K}]\!]$ , which is of the same dimension as  $\mathcal{Y}$  but its entries are continuous-valued, admits a low-rank CP structure. The goal is to estimate  $\Theta$  from the large binary tensor  $\mathcal{Y}$  of arbitrary order  $K$ .

The low-rank Bernoulli model (1) was first proposed in the context of matrices (Collins et al., 2002; De Leeuw, 2006; Lee et al., 2010) and recently extended to higher-order tensors (Mažgut et al., 2014; Hong et al., 2018). We are particularly interested in the modern high-dimensional setting where the number of effective parameter increases with sample size. As the data tensor grows in dimension, it is important to investigate the trade-offs between statistical and computational accuracy. Such consideration leads to the fundamental inquiry into the estimation problem, such as the signal-to-noise ratio for which a reliable estimation of  $\Theta$  is possible, the intrinsic hardness (in terms of minimax rate) of the binary problem compare to its continuous-valued counterpart, and the estimation properties specific to a particular algorithm vs. to all algorithms. From a practical standpoint, understanding these questions can reveal gaps between a current algorithm compared to the statistical optimality, and in turn guides the algorithm design to achieve the best trade-offs. One of our contribution is to provide answers to these questions.

We would like to emphasize that there are two types of properties that an estimator may possess. The first type is the algorithm-dependent property which quantifies the impact of algorithm implementation (such as choice of loss function, initialization, iterations, etc) on the output. The second type is the algorithm-independent or statistical property which characterizes the population-wide information one can obtain from a particular data instance. Earlier development of binary tensor decomposition (Miettinen, 2011; Mažgut et al., 2014; Rai et al., 2015; Hong et al., 2018) focuses on the algorithm effectiveness, but

does not address the population properties of the estimation problem. As we shall see, the results in this paper fill the gap by addressing both types.

**Related work.** Our work is closely related to but also clearly distinctive from several lines of existing research. We survey the main ideas that have emerged.

*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  $[0, 1]$ . As we will see in Section 2.1, model (1) can be regarded 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 generalized linear model (GLM). On one hand, the GLM enjoys invariance properties; that is, reversing the entry coding of  $\mathcal{Y}$  only changes the sign, but not the low-rank structure, of the parameter  $\Theta$ . On the other hand, the binary problem exhibits a distinctive “dithering” effect (Davenport et al., 2014) which necessitates the presence of stochastic noise in order to estimate  $\Theta$ . Both phenomena are clearly contrary to the behavior of continuous-valued tensor decomposition. We will investigate them in more detail in Section 3.4.

*Binary matrix decomposition.* When  $K = 2$ , the Bernoulli tensor model (1) reduces to binary or logit principal component analysis (PCA) (Collins et al., 2002; De Leeuw, 2006; Lee et al., 2010). While tensors are conceptual generalization of matrices, it has been shown that higher-order tensor problems possess fundamental challenges compared to matrices (Kolda and Bader, 2009). Note that, for the matrix case, the rank  $R \leq \min(d_1, d_2)$ , and the factor matrices 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 (Kruskal, 1977; Bhaskara et al., 2014). Unlike matrices, tensors do not always admit orthogonal decomposition, and the tensor rank  $R$  can exceed the dimension. These different properties make the earlier algorithms that built upon matrix decomposition inapplicable to tensors.

*1-bit completion.* Our work also has connection 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 3.2. 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) 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 deal with the binary setting as we

do, but they take a model-free approach to *approximate* the data instance. One important difference between the data-driven approach and our proposal is that we focus on *parameter estimation* in a population model. The population interpretation offers new insight on the effectiveness of dimension reduction. Having a population model allows one to tease apart the algorithmic error vs. statistical error. In this respect, our proposal is orthogonal to the boolean tensor decomposition and other existing data mining techniques (Erdos and Miettinen, 2013a; Hu et al., 2018). We note that, in practice, it is usually unknown which approach would perform the best for the dataset at hand; this is essentially a model selection problem. We assess the empirical robustness of both approaches in Section 5.

*Bayesian binary tensor decomposition.* Last but not least, there have been a number of Bayesian binary tensor decomposition algorithms (Nickel et al., 2011; Rai et al., 2014, 2015). Most of those works 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 reveal some algorithm-independent properties of the problem, such as the SNR phase diagram and minimax rate. As such, our results provide a unified framework for the aforementioned works. Besides, we provide a frequentist-type solution which is computationally more tractable than a Bayesian solution.

**Main contributions.** We summarize our main results that set our work aside from the existing literature.

First, we develop a framework to quantify the hardness of the binary tensor problem. It is shown that the Bernoulli tensor model is equivalent to entrywise quantization of a latent noisy continuous-valued tensor. We characterize the impact of latent signal-to-noise ratio (SNR) on the tensor recovery accuracy, as we identify three different phases for tensor recovery according to SNR; see Table 1 in Section 3.4. 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 is 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. We handle this challenge via a constrained estimation (see Section 3.4).

Second, we propose a new estimation method for binary tensor decomposition and establish its statistical properties. Note that in our problem the number of parameter increases with dimension, so the classical MLE theory does not directly apply. We leverage the recent development in random tensor theory and high-dimensional statistics to establish the error bound 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,  $d_1 = \dots = d_K = d$ , we obtain a convergence rate  $\asymp d^{-K+1}$  for estimating  $\Theta$ . This result outperforms the rate from “best matrixization” ( $\asymp d^{-\lfloor K/2 \rfloor}$ , where  $\lfloor K/2 \rfloor$  is the integer part of  $K/2$ ; see Section 3.2) and 1-bit tensor completion ( $\asymp d^{-(K-1)/4}$ ; see Ghadermarzy et al. (2018)). To our knowledge, these statistical guarantees are among the first for binary tensor observations.

Lastly, we supplement the above general theory by proposing a block relaxation algorithm. We investigate the landscape of objective function and establish the locally linear

convergence of the optimization. The main assumption of convergence requires the local uniqueness (modulo scaling) for the CP decomposition of  $\Theta$ , which is mild for tensors of order 3 or higher (Kruskal, 1977; Bhaskara et al., 2014). The algorithm-dependent error bound reveals an interesting interaction between the computational efficiency and the statistical convergence. We illustrate the efficacy of our algorithm through simulations and data applications.

**Notation and organization.** We adopt the following notation throughout the article. We use  $\mathcal{Y} = \llbracket y_{i_1, \dots, i_K} \rrbracket \in \mathbb{F}^{d_1 \times \dots \times d_K}$  to denote an order- $K$  ( $d_1, \dots, d_K$ )-dimensional tensor over a field  $\mathbb{F}$ . 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  $\|\mathcal{Y}\|_F = (\sum_{i_1, \dots, i_K} y_{i_1, \dots, i_K}^2)^{1/2}$ , and the infinity norm of  $\mathcal{Y}$  is defined as  $\|\mathcal{Y}\|_\infty = \max_{i_1, \dots, i_K} |y_{i_1, \dots, i_K}|$ . We use capital letters,  $\Theta, \mathcal{Y}, \mathbf{A}$  etc, to denote tensors and matrix, and use little letters,  $\theta, \mathbf{a}$  etc, to denote scales and vectors. We use  $\mathbf{a} \otimes \mathbf{b}$  to denote the kronecker product of vectors  $\mathbf{a}$  and  $\mathbf{b}$ , and  $\mathbf{A} \odot \mathbf{B}$  to denote the Khatri-Rao product of matrices  $\mathbf{A}$  and  $\mathbf{B}$ . We use  $\mathbf{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$  to denote the  $(d - 1)$ -dimensional unit sphere, and the shorthand  $n$ -set  $\{1, \dots, n\}$  to denote  $n \in \mathbb{N}_+$ .

The rest of the article is organized as follows. We describe the low-rank Bernoulli tensor model and its connection with 1-bit observation model in Section 2. We develop a constrained likelihood-based estimation procedure in Section 3. Statistical properties are presented, including the upper bound, the minimax lower bound, and the phase-transition on the tensor recovery accuracy. In Section 4, we develop an alternating updating algorithm with convergence guarantees. 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 Appendix.

## 2. Model

### 2.1 Low-rank Bernoulli model

For the binary tensor  $\mathcal{Y} = \llbracket y_{i_1, \dots, i_K} \rrbracket \in \{0, 1\}^{d_1 \times \dots \times d_K}$ , we assume its entries are realizations of independent Bernoulli random variables, in that

$$\mathcal{Y} | \Theta \sim \text{Bernoulli } \{f(\Theta)\}, \quad \text{with } \mathbb{P}(y_{i_1, \dots, i_K} = 1) = f(\theta_{i_1, \dots, i_K}), \quad (2)$$

for all  $(i_1, \dots, i_K) \in [d_1] \times \dots \times [d_K]$ . In this model,  $f: \mathbb{R} \rightarrow [0, 1]$  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 GLM language,  $f$  is often referred to as the “inverse link function”. When no confusion arises, we also call  $f$  the “link function”. The parameter tensor  $\Theta = \llbracket \theta_{i_1, \dots, i_K} \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$  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  $\Theta$ . Note that this assumption does not rule out the marginal correlation among entries in  $\mathcal{Y}$ . The conditional independence assumption is again commonly adopted in the literature (Collins et al., 2002; De Leeuw, 2006; Lee et al., 2010).

Furthermore, we assume the parameter tensor  $\Theta$  admits a rank- $R$  CP decomposition,

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

where  $\lambda_r \in \mathbb{R}_+$ , with  $\lambda_1 \geq \dots \geq \lambda_K$  without loss of generality,  $\mathbf{a}_r^{(k)} \in \mathbf{S}^{d_k-1}$ , for  $r \in [R], k \in [K]$ , and  $\Theta$  cannot be written as a sum of fewer than  $R$  outer products. This low-rank structure in (3) dramatically reduces the number of parameters in  $\Theta$ , 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). This structure is an effective and frequently used tool in high-dimensional tensor 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  $\Theta$  is small, is independent of the tensor dimension, and offers a reasonable approximation to the truth (Zhou et al., 2013; Bhaskar and Javanmard, 2015).

Combining (2) and (3) leads to our low-rank Bernoulli model. We seek to estimate the rank- $R$  tensor  $\Theta$  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  $\Theta$  from a noisy observation  $\mathcal{Y} = \Theta + \mathcal{E}$ , where  $\mathcal{E} \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_K}$  collects i.i.d. Gaussian noises. In this regard, our proposal is a generalization that is analogous to the one from a Gaussian linear model to a generalized linear model (GLM). In next section, we will see a close connection between a real-valued problem and a binary problem.

## 2.2 Latent variable model interpretation

Before turning to estimation, we 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  $\Theta = [\theta_{i_1, \dots, i_K}] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$  with a rank- $R$  CP structure. Suppose that we cannot not directly observe  $\Theta$ . Instead, we observe the quantized version  $\mathcal{Y} = [y_{i_1, \dots, i_K}] \in \{0, 1\}^{d_1 \times \cdots \times d_K}$  following the scheme

$$y_{i_1, \dots, i_K} = \begin{cases} 1 & \text{if } \theta_{i_1, \dots, i_K} + \varepsilon_{i_1, \dots, i_K} \geq 0, \\ 0 & \text{if } \theta_{i_1, \dots, i_K} + \varepsilon_{i_1, \dots, i_K} < 0, \end{cases} \quad (4)$$

where  $\mathcal{E} = [\varepsilon_{i_1, \dots, i_K}]$  is a noise tensor. The binary tensor is generated from  $\mathcal{Y} = \text{sign}(\Theta + \mathcal{E})$ , and associated latent tensor is  $\Theta + \mathcal{E}$ . Here we allow the sign function  $\text{sign}(x) \stackrel{\text{def}}{=} \mathbb{1}_{\{x \geq 0\}}$  to

be applied to tensors in an element-wise manner. From the viewpoint of (4), the tensor  $\Theta$  is not merely an argument in the Bernoulli tensor model (c.f. Equation (2)); the tensor  $\Theta$  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  $\mathbb{P}(\varepsilon \leq \theta) = 1 - f(-\theta)$ , 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 an 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  $f(\theta) = (1 + e^{-\theta/\sigma})^{-1}$  and the scale parameter  $\sigma > 0$ . Equivalently, the noise  $\varepsilon_{i_1, \dots, i_K}$  in (4) follows i.i.d. logistic distribution with the scale parameter  $\sigma$ .

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

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

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

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

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

### 3. Statistical properties

#### 3.1 Rank-constrained likelihood-based estimation

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

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

where  $q_{i_1, \dots, i_K} = 2y_{i_1, \dots, i_K} - 1$  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:

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

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  $\Theta_{\text{true}}$ . The candidate tensor of our interest satisfies two constraints. The first is that  $\Theta$  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  $\Theta$  are bounded in absolute value by a constant  $\alpha \in \mathbb{R}_+$ . We refer to  $\alpha$  as the “signal” bound of  $\Theta$ . This infinity-norm condition is a technical assumption to aid the recovery of  $\Theta$  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).

Let  $\hat{\Theta}_{\text{MLE}} = \arg \max_{\Theta \in \mathcal{D}} \mathcal{L}_Y(\Theta)$  in (5), and we call it the maximum likelihood estimator as it maximizes the log-likelihood function. Note that classical MLE theory does not directly apply here because the number of parameters  $p_e$  grows with the sample size  $\{d_k\}$ . The estimation accuracy is inevitably woven around three quantities: tensor dimension, rank, and signal-to-noise ratio. In the next section, we will establish the performance bound for  $\hat{\Theta}_{\text{MLE}}$ .

### 3.2 Performance upper bound

We define two quantities  $L_\alpha$  and  $\gamma_\alpha$  to control the “steepness” and “convexity” of the link function  $f$ . Let

$$L_\alpha = \sup_{|\theta| \leq \alpha} \left\{ \frac{\dot{f}(\theta)}{f(\theta)(1-f(\theta))} \right\}, \quad \text{and} \quad \gamma_\alpha = \inf_{|\theta| \leq \alpha} \left( \frac{\dot{f}^2(\theta)}{f^2(\theta)} - \frac{\ddot{f}(\theta)}{f(\theta)} \right), \quad (6)$$

where  $\dot{f}(\theta) = df(\theta)/d\theta$ , and  $\alpha$  is the bound on the entry-wise infinity-norm of  $\Theta$ . When  $\alpha$  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

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

We assess the estimation accuracy using the deviation in the Frobenius norm. For the true coefficient tensor  $\Theta_{\text{true}} \in \mathbb{R}^{d_1 \times \dots \times d_K}$  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  $\mathcal{Y} \in \{0, 1\}^{d_1 \times \dots \times d_K}$  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  $1 - \exp(-C_1 \log K \sum_k d_k)$ ,

$$\text{Loss}(\hat{\Theta}_{MLE}, \Theta_{true}) \leq \min \left\{ 2\alpha, \frac{C_2 L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}} \right\}. \quad (7)$$

The proof of the Theorem 1 is given in Section B.1. 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 (7). A matching lower bound will be provided in Section 3.3.

To gain insight into this error bound, we consider a special setting where the dimensions are the same in all modes, i.e.,  $d_1 = \dots = d_K = d$ . In such a case, our bound (7) reduces to

$$\text{Loss}(\hat{\Theta}_{MLE}, \Theta_{true}) \leq O\left(\frac{1}{d^{(K-1)/2}}\right), \quad (8)$$

for a fixed rank  $R$  and signal bound  $\alpha$  as  $d \rightarrow \infty$ . The proposed estimator achieves consistency with inverse-polynomial convergence rate. Compared to previous literature on error bound for 1-bit tensor recovery (Ghadermarzy et al., 2018)

$$\text{Loss}(\hat{\Theta}, \Theta_{true}) \leq O\left(\frac{1}{d^{(K-1)/4}}\right), \quad (9)$$

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  $\Theta$ , 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  $d \rightarrow \infty$ .

As an immediate corollary of Theorem 1, we obtain the explicit upper bound (7) 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  $1 - \exp(-C' \log K \sum_k d_k)$ ,

$$\text{Loss}(\hat{\Theta}_{MLE}, \Theta_{true}) \leq \min \left\{ 2\alpha, C(\sigma, \alpha) \sqrt{\frac{R^{K-1} K \sum_k d_k}{\prod_k d_k}} \right\}, \quad (10)$$

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

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

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  $\alpha$  and the noise level  $\sigma$  will be investigated in Section 3.4.

### 3.3 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 information theory and is thus 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  $\hat{\Theta}_{\text{MLE}}$ . The second lower bound is for any estimator  $\tilde{\Theta}$  from the “unquantized” version of the continuous-valued tensor decomposition. This allows us to evaluate how much information is lost by quantizing a continuous-valued tensor to a binary one (see Model (4)). With a little abuse of notation, we use  $\mathcal{D}(R, \alpha)$  to denote the set of tensors with rank bounded by  $R$  and infinity norm bounded by  $\alpha$ .

The next theorem establishes the information-theoretical lower bound for any estimator  $\hat{\Theta}$  in  $\mathcal{D}(R, \alpha)$  under model (2).

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

$$\inf_{\hat{\Theta}} \sup_{\Theta_{\text{true}} \in \mathcal{D}(R, \alpha)} \mathbb{P} \left\{ \text{Loss}(\hat{\Theta}, \Theta_{\text{true}}) \geq c_0 \min \left( \alpha, \sigma \sqrt{\frac{R d_{\max}}{\prod_k d_k}} \right) \right\} \geq \beta_0. \quad (11)$$

This proof of this theorem is based on the information-theoretical arguments, and is given in Section B.2. 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. Note that in Theorem 1, we do not place any constraint on the rank  $R$ .

We now compare the lower bound (11) to the upper bound (10), as the tensor dimension  $d_k \rightarrow \infty$  while the signal bound  $\alpha$  and the noise level  $\sigma$  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,  $d_1 = \dots = d_K = d$ , the convergence rate for estimating  $\Theta$  using our tensor method is  $O(d^{-K+1})$ , while the “best” unfolding solution that unfolds a tensor into a near-square matrix (Mu et al., 2014) gives the convergence rate  $O(d^{-\lfloor K/2 \rfloor})$ , where  $\lfloor K/2 \rfloor$  is the integer part of  $K/2$ . This gap

between the two rates highlights the importance of binary decomposition that specifically takes advantage of the multi-mode structure in tensors.

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  $\mathcal{Y} \in \mathbb{R}^{d_1 \times \dots \times d_K}$  is an order- $K$  continuous-valued tensor from the model  $\mathcal{Y} = \Theta_{true} + \mathcal{E} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ , 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  $\Theta_{true}$ . Then there exist absolute constants  $\beta_0 \in (0, 1)$  and  $c_0 > 0$  such that*

$$\inf_{\tilde{\Theta}} \sup_{\Theta_{true} \in \mathcal{D}(R, \alpha)} \mathbb{P} \left\{ Loss(\tilde{\Theta}, \Theta_{true}) \geq c_0 \min \left( \alpha, \sigma \sqrt{\frac{R d_{\max}}{\prod_k d_k}} \right) \right\} \geq \beta_0. \quad (12)$$

The proof of Theorem 3 is given in Section B.3. This lower bound (12) 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.

### 3.4 Phase diagram

The error bounds we have established depend on the signal bound  $\alpha$  and the noise level  $\sigma$ . 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 phases under the case when  $d_1 = \dots = d_K = d$ . 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 (10) 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  $\sigma$  goes to zero, the problem essentially reverts to the noiseless case where an accurate estimation of  $\Theta$  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,  $\Theta_1 = \mathbf{a}_1 \otimes \mathbf{a}_2 \otimes \mathbf{a}_3$  and  $\Theta_2 = \text{sign}(\mathbf{a}_1) \otimes \text{sign}(\mathbf{a}_2) \otimes \text{sign}(\mathbf{a}_3)$ , would lead to the same observation  $\mathcal{Y}$  in the absence of noise, and thus recovery of  $\Theta$  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	$\sigma e^{\alpha^2/\sigma^2} d^{-(K-1)/2}$	$\sigma d^{-(K-1)/2}$	$\alpha$
Continuous	$\sigma d^{-(K-1)/2}$	$\sigma d^{-(K-1)/2}$	$\alpha$

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

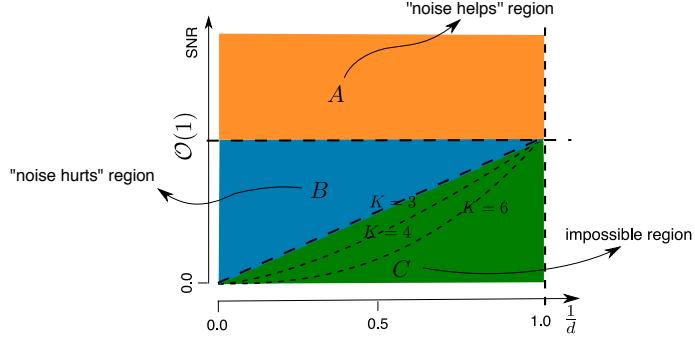


Figure 1: Phase diagram with respect to SNR. (A) “Noise helps” region (in yellow): error decreases with noise . (B) “Noise hurts” region (in blue): error increases with noise. (C) Impossible region (in green): A consistent estimator of  $\Theta$  is impossible. The dashed line between regions (B) and (C) depicts the boundary  $d^{-(K-1)/2}$  as  $K$  varies. Note that the origin in the  $x$ -axis corresponds to the high dimension region ( $d^{-(K-1)/2} \rightarrow 0$ ) which is of our 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 (10) scales linearly with  $\sigma$ . Comparing the lower bound (12) when estimating the signal from the unquantized tensor, to the upper bound (10) 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  $\Theta$  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  $\Theta$  becomes impossible. In this regime, a trivial zero estimator achieves the minimax rate.

## 4. Implementation

### 4.1 Block relaxation algorithm

Earlier we have presented the statistical properties of the tensor estimation. We now introduce an algorithm solving (5). For notational convenience, we drop the subscript  $\mathcal{Y}$  in  $\mathcal{L}_{\mathcal{Y}}(\Theta)$  and simple write  $\mathcal{L}(\Theta)$ . The objective function  $\mathcal{L}(\Theta)$  is concave in  $\Theta$  when the

link function  $f$  is log-concave in  $\theta_{i_1, \dots, i_K}$ . 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 for each block of parameters while fixing the others.

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

$$\mathbf{A}_k = [\mathbf{a}_1^{(k)}, \dots, \mathbf{a}_R^{(k)}] \in \mathbb{R}^{d_k \times R}, k \in [K-1], \text{ and } \mathbf{A}_K = [\lambda_1 \mathbf{a}_1^{(K)}, \dots, \lambda_R \mathbf{a}_R^{(K)}] \in \mathbb{R}^{d_K \times R}, \quad (13)$$

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

$$\max_{\mathbf{A}} \mathcal{L}(\Theta(\mathbf{A})), \quad \text{subject to } \Theta(\mathbf{A}) \in \mathcal{D}. \quad (14)$$

Although the objective function in (14) 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  $\mathbf{A}_k^{(t)}$  denote the  $k$ th factor matrix at the  $t$ th iteration, and

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

Let  $\mathcal{Y}(:, j(k), :)$  denote the sub-tensor of  $\mathcal{Y}$  at the  $j$ th position of the  $k$ th mode,  $j = 1, \dots, d_k$ ,  $k = 1, \dots, K$ , and  $\text{vec}(\cdot)$  is the operator that turns a tensor into a vector. Then the update  $\mathbf{A}_k^{(t+1)}$  can be obtained row-by-row by solving  $d_k$  separate GLMs, where each GLM takes  $\text{vec}\{\mathcal{Y}(:, j(k), :)\} \in \mathbb{R}^{(\prod_{i \neq k} d_i) \times 1}$  as the “response”,  $\mathbf{A}_{-k}^{(t)} \in \mathbb{R}^{(\prod_{i \neq k} d_i) \times R}$  as the “predictors”, and the  $j$ th row of  $\mathbf{A}_k$  as the “regression coefficient”,  $j = 1, \dots, d_k$ ,  $k = 1, \dots, K$ . 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  $\{\mathbf{A}_k^{(t+1)}\}$  by performing a line search,

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

We then update  $\mathbf{A}_k^{(t+1)} = \gamma^* \mathbf{A}_k^{(t)} + (1 - \gamma^*) \mathbf{A}_k^{(t+1)}$ , and normalize the columns of  $\mathbf{A}_k^{(t+1)}$ . 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  $\mathbf{A}^{(t)}$  and  $\Theta^{(t)} = \Theta(\mathbf{A}^{(t)})$ . To simplify the analysis, we set the hyper-parameter  $\alpha$  to infinity, which essentially poses no prior on the tensor magnitude. We make the following assumptions:

**Algorithm 1** Binary tensor decomposition

**Input:** Binary tensor  $\mathcal{Y} \in \{0, 1\}^{d_1 \times \dots \times d_K}$ , link function  $f$ , rank  $R$ , and entrywise bound  $\alpha$ .

**Output:** Rank- $R$  coefficient tensor  $\Theta$ , along with the factor matrices  $\mathbf{A} = (\mathbf{A}_1, \dots, \mathbf{A}_K)$ .

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

- (A1) (Regularity) The log-likelihood  $\mathcal{L}(\mathbf{A})$  is continuous and the set  $\{\mathbf{A} : \mathcal{L}(\mathbf{A}) \geq \mathcal{L}(\mathbf{A}^{(0)})\}$  is compact.
- (A2) (Strictly local maximum) 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 Hessian is non-singular at the solution.
- (A3) (Local uniqueness) The set of stationary points of  $\mathcal{L}(\mathbf{A})$  are isolated (modulo scaling).
- (A4) (Local Lipschitz) The tensor rank- $R$  representation  $\Theta = \Theta(\mathbf{A})$  is called locally Lipschitz at  $\mathbf{A}^*$ , if there exists two constants  $c_1, c_2 > 0$  such that

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

for  $\mathbf{A}', \mathbf{A}''$  sufficiently close to  $\mathbf{A}^*$ . Here  $\mathbf{A}', \mathbf{A}''$  represent the block variables subject to convention (13).

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  $\mathbf{A}_k$ . Note that the full Hessian needs not to be negative definite in all variables simultaneously. We consider this as a reasonably mild assumption, as similar conditions have often been imposed in many non-convex problems (Uschmajew, 2012; Zhou et al., 2013; Sun et al., 2017). Assumptions (A2)–(A4) guarantee the (local) uniqueness of the CP decomposition  $\Theta = \Theta(\mathbf{A})$ . The conditions exclude the case of rank-degeneracy (e.g., if the tensor  $\Theta$  can be written in fewer than  $R$  factors, or if the columns of  $\mathbf{A}_{-k}^{(t)}$  are linearly dependent in the GLM update) and the case of non-unique decompositions (e.g., if the decomposition  $\Theta$  can be smoothly changed beyond scaling of the factors.) Again, these requirements are fairly mild for tensors of order 3 or higher. For more discussions on decomposition uniqueness and its implication on the optimization, we refer to Kruskal (1977); Uschmajew (2012); Zhou et al. (2013).

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

- (i) (Global convergence) Any sequence  $\mathbf{A}^{(t)} = (\mathbf{A}_1^{(t)}, \dots, \mathbf{A}_K^{(t)})$  generated by Algorithm 1 converges to a stationary point of  $\mathcal{L}(\mathbf{A})$ .
- (ii) (Locally linear convergence) Let  $\mathbf{A}^*$  be a local maximizer of  $\mathcal{L}$ . There exists an  $\varepsilon$ -neighborhood of  $\mathbf{A}^*$ , such that for any starting point  $\mathbf{A}^{(0)}$  in this neighborhood, the iterates  $\mathbf{A}^{(t)}$  of Algorithm 1 linearly converge to  $\mathbf{A}^*$ ,

$$\|\mathbf{A}^{(t)} - \mathbf{A}^*\|_F \leq \rho^t \|\mathbf{A}^{(0)} - \mathbf{A}^*\|_F, \quad (15)$$

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

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

Proposition 1(ii) reveals that every local maximizer of  $\mathcal{L}$  is an attractor of Algorithm 1. This is a nice property ensuring the exponential decay of the error near a local maximum. To appreciate the merit of tensor decomposition over matrices, we show that the same proposition often fails for matrices. Let us consider an example in a 2-by-2 matrix problem. Suppose that the local maximizer is  $\Theta^* = \Theta^*(\mathbf{e}_1, \mathbf{e}_2) = \mathbf{e}_1^{\otimes 2} + \mathbf{e}_2^{\otimes 2}$ , where  $\mathbf{e}_1, \mathbf{e}_2$  are canonical vectors in  $\mathbb{R}^2$ . There is no region of attraction near the block variable  $\mathbf{A}^* = (\mathbf{e}_1, \mathbf{e}_2)$ . In fact, one can construct a point  $\mathbf{A}^{(0)} = (\mathbf{a}_1, \mathbf{a}_2)$  with  $\mathbf{a}_1 = (\sin \theta, \cos \theta)'$  and  $\mathbf{a}_2 = (\cos \theta, -\sin \theta)'$ . The point  $\mathbf{A}^{(0)}$  can be made arbitrarily close to  $\mathbf{A}^*$  (via tuning  $\theta$ ) but the algorithm iteration initialized from  $\mathbf{A}^{(0)}$  will never converge to  $\mathbf{A}^*$ . This behavior occurs when the matrix has degenerate singular-values. In contrast, a 2-by-2-by-2 tensor problem with maximizer  $\tilde{\Theta}^* = \tilde{\Theta}^*(\mathbf{e}_1, \mathbf{e}_2) = \mathbf{e}_1^{\otimes 3} + \mathbf{e}_2^{\otimes 3}$  possesses locally unique decomposition (and thus isolated stationary points). The block variable  $\mathbf{A}^*$  exhibits a basin of attraction with respect to the tensor objective.

Combining Proposition 1 and Theorem 1, we achieve the following results on the interplay between statistical and computational errors.

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

$$Loss(\Theta(\mathbf{A}^{(t)}), \Theta_{true}) \leq \underbrace{C_1 \rho^t \varepsilon}_{\text{algorithmic error}} + \underbrace{\frac{C_2 L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}}}_{\text{statistical error}}, \quad (16)$$

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

Theorem 4 provides the estimation error at each iteration of our Algorithm 1. The bound (16) consists of two terms. The first term is the algorithmic error, and the second term is the statistical error. The algorithmic error decays exponentially with the iteration, whereas the statistical error remains the same as  $t$  grows. The statistical error is unavoidable because it reflects the fundamental error due to estimating from a noisy tensor (see Theorem 2).

The result reveals an interesting interplay between the computational and statistical efficiency. For tensors with equal dimension  $d_1 = \dots = d_K = d$ , when the iteration number satisfies

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

the computational error is dominated by the statistical error. The bound provides a meaningful guidance for the needed number of iterations  $T$  to reach a balance between the two considerations.

### 4.3 Missing data and tuning parameters

When some tensor entries  $y_{i_1, \dots, i_K}$  are missing, we replace the objective function  $\mathcal{L}_Y(\Theta)$  with  $\sum_{(i_1, \dots, i_K) \in \Omega} \log f(q_{i_1, \dots, i_K} \theta_{i_1, \dots, i_K})$ , where  $\Omega \subset [d_1] \times \dots \times [d_K]$  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  $y_{i_1, \dots, i_K}$  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.

Our procedure provides a straightforward way to handle missing data. As a by-product, our tensor decomposition output can also be used for missing value prediction. That is, we predict the missing values  $Y_{i_1, \dots, i_K}$  using  $f(\hat{\Theta}_{i_1, \dots, i_K})$ , where  $\hat{\Theta}$  is the low-rank 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  $\mathbb{P}(Y_{i_1, \dots, i_K} = 1)$ .

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

$$\hat{R} = \min_{R \in \mathbb{R}_+} \text{BIC}(R) = \min_{R \in \mathbb{R}_+} \left\{ -2\mathcal{L}_Y\{\hat{\Theta}(R)\} + p_e(R) \log \left( \prod_k d_k \right) \right\}, \quad (17)$$

where  $\hat{\Theta}(R)$  is the estimated coefficient tensor  $\Theta$  under the working rank  $R$ , and  $p_e(R)$  is the effective number of parameters. 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,

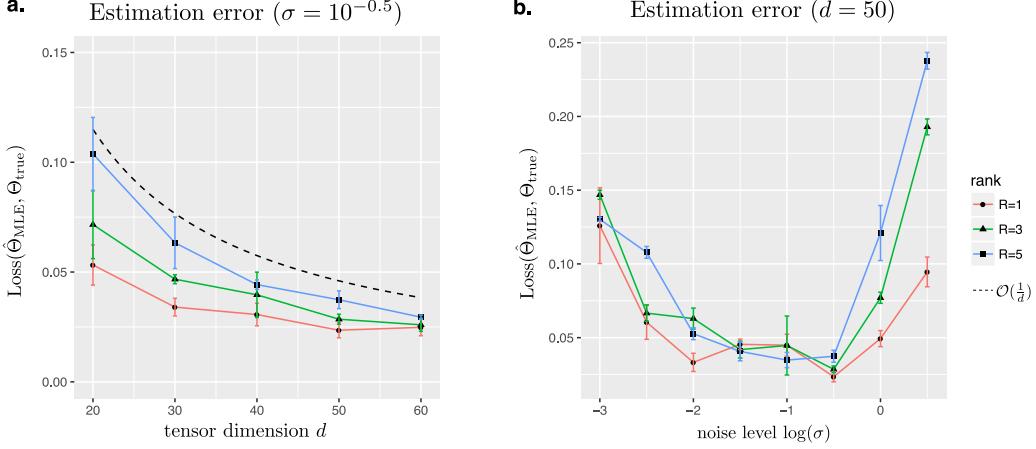


Figure 2: Estimation error of binary tensor decomposition. (a) Estimation error as a function of the tensor dimension  $d = d_1 = d_2 = d_3$ . (b) Estimation error as a function of the noise level.

and this complexity matches the classical continuous-valued tensor decomposition. More precisely, the update of  $\mathbf{A}_k$  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 the earlier sections we have provided the asymptotical guarantees of our binary tensor method in the high-dimensional region. In this section we investigate the finite-sample performance under 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  $\Theta_{\text{true}} = \sum_{r=1}^R \mathbf{a}_r^{(1)} \times \mathbf{a}_r^{(2)} \times \mathbf{a}_r^{(3)}$ , and the entries of  $\mathbf{a}_r^k$  are i.i.d. drawn from  $\text{Unif}[-1,1]$ . Without loss of generality, we scale  $\Theta_{\text{true}}$  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  $d \in \{20, 30, \dots, 60\}$ , and the noise level  $\sigma \in \{10^{-3}, 10^{-2.5}, \dots, 10^{0.5}\}$ . Logistic link function is employed, and we report the estimation error averaged across  $n_{\text{sim}} = 30$  replications.

Figure 2a 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 2b plots the estimation error as a function of the noise level  $\sigma$  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 (see Section 3.4 and Figure 1).

We assess the effectiveness of using BIC to select the tensor rank. We consider the tensor dimension  $d \in \{20, 40, 60\}$  and rank  $R \in \{5, 10, 20, 40\}$ . There are several scenarios in which rank exceeds the dimension; this allows us to investigate over-complete tensor decomposition. 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  $n_{\text{sim}} = 30$  data tensors under the Bernoulli tensor model (2). We minimize the 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. Although the block relaxation in Algorithm 1 has no theoretical guarantee to land to the global optimum, in practice, we often find that the converged points have satisfactory objective values. Based on Theorem 4 and Remark 2 in the Appendix, a sufficient (but not necessary) condition for  $\hat{\Theta}$  to have the desired convergence rate is that  $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}) \geq \mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}})$ . Such conditions are met for all of  $n_{\text{sim}} = 30$  replications in our experiment. Figure 3 shows 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 observed that the algorithm generally converges quickly upon random initializations, and it usually takes fewer than 8 steps for the relative decrease in the objective to be below 3% even for a large  $d$  and  $R$ . The average computation time per iteration is shown in the legend. For instance, when  $\prod_i d_i = 60^3$  and  $R = 10$ , each iteration of Algorithm 1 on average takes fewer than 3 seconds.

True rank	$\sigma = 0.1$			$\sigma = 0.01$		
	$d = 20$	$d = 40$	$d = 60$	$d = 20$	$d = 40$	$d = 60$
$R = 5$	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.

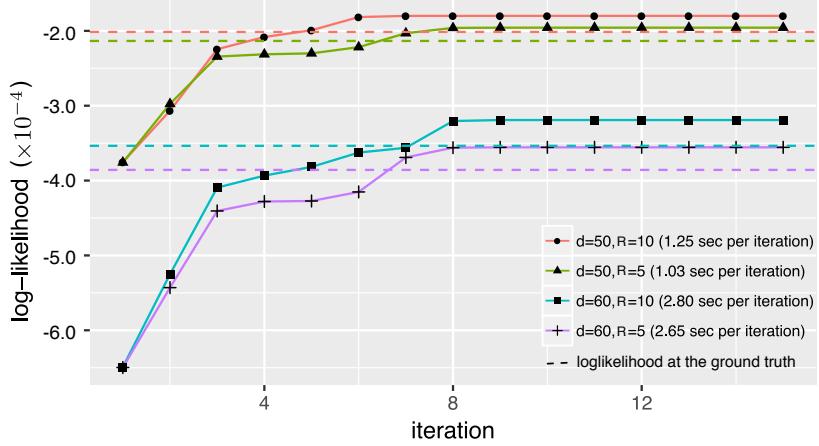


Figure 3: Trajectory of the objective function over iterations with varying  $d$  and  $R$ .

## 5.2 Comparison with other methods

We next compare our method (BTF\_Alternating<sup>1</sup>) with the following methods:

- Boolean tensor factorization (BooleanTF) (Miettinen, 2011; Erdos and Miettinen, 2013b; Rukat et al., 2018). BooleanTF decomposes a binary tensor into binary factors. It aims to cover 0’s and 1’s based on a set of logic rules among factors. We implement the software from Rukat et al. (2018).
- Bayesian tensor factorization (BTF\_Bayesian) (Rai et al., 2014). BTF\_Bayesian uses Expectation Maximization (EM) algorithm to decompose a binary tensor into continuous-valued factors. The method imposes Gaussian prior on the factor entries and multiplicative gamma process prior on the factor weights  $\{\lambda_r\}$ .
- Bernoulli tensor factorization with gradient descent (BTF\_Gradient) (Hong et al., 2018). BTF\_Gradient uses gradient descent algorithm to decompose a binary tensor into continuous-valued factors. We implement the development version of the algorithm which is an upcoming part of the toolbox in Matlab.

The four methods under comparison fall into two categories. BooleanTF differs from the other three in both the cost function and the output formats. We include this method for the purpose of assessing model misspecification. The other three methods are based on Bernoulli models (2) but with different implementations. In particular, TF\_Bayesian utilizes Bayesian optimization whereas the other two utilize frequentist optimization.

We report two metrics for error comparison. The first one is root mean square error (RMSE), defined as  $\text{RMSE} = \frac{1}{\sqrt{\prod_k d_k}} \|\widehat{\mathbb{E}(\mathcal{Y})} - \mathbb{E}(\mathcal{Y})\|_F$ , where  $\widehat{\mathbb{E}(\mathcal{Y})}$  denotes the estimated success probability tensor. For BooleanTF, this quantity is represented as posterior mean of  $\mathcal{Y}$  (Miettinen, 2011), and for the other three BTF methods, we take  $\widehat{\mathbb{E}(\mathcal{Y})} = \text{logit}(\hat{\Theta})$ . The second metric is misclassification error rate (MER), defined as  $\text{MER} = \frac{1}{\sqrt{\prod_k d_k}} |\mathbf{1}_{\widehat{\mathbb{E}(\mathcal{Y})} \geq 0.5} -$

1. Software implementation: <https://github.com/Miaoyanwang/Binary-Tensor>

$\mathbb{1}_{\mathbb{E}(\mathcal{Y}) \geq 0.5}$ . Here we allow the indicator function to be applied to tensors in an element-wise manner. The RMSE and MER represent two aspects of 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 binary tensors from a range of misspecified models which are detailed in the next section. In each class, the signal tensors do not always follow the CP structure exactly. This allows us to evaluate the robustness of our method under potential model misspecification.

### 5.2.1 LOGIT BOOLEAN MODEL

We first consider the logic boolean tensors. The simulation set-up is similar to [Rukat et al. \(2018\)](#). We first simulate noiseless tensors  $\mathcal{Y} = [\![y_{ijk}]\!]$  from the following model,

$$y_{ijk} = \bigvee_{r=1}^R \bigwedge_{ijk} a_{ir} b_{jr} c_{kr}, \quad (18)$$

$$\text{with } a_{ir} \sim \text{Ber}(p_{ir}^a), b_{jr} \sim \text{Ber}(p_{jr}^b), c_{kr} \sim \text{Ber}(p_{kr}^c), \quad (19)$$

where the binary factor entries  $\{a_{ir}\}$ ,  $\{b_{jr}\}$ ,  $\{c_{kr}\}$  are mutually independent with each other, 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

$$\mathbb{E}(y_{ijk}) = 1 - \prod_{r=1}^R \left(1 - p_{ir}^a p_{jr}^b p_{kr}^c\right).$$

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 tensors with  $d_1 = d_2 = d_3 = 50$  and boolean rank  $R \in \{10, 15, 20, 25, 30\}$ . The factor probabilities  $\{p_{ir}^a\}$ ,  $\{p_{jr}^b\}$ ,  $\{p_{kr}^c\}$  are generated i.i.d. from Beta(2,4).

We apply the four methods to  $n_{\text{sim}} = 20$  replicates in each scenario. In all methods, we use default parameters and select  $R$  using their recommended selection functions. For our method BTF\_Alternating, we set the hyper-parameter  $\alpha$  to infinity, which essentially poses no prior on the tensor magnitude. Note that the definition of boolean rank is different from the rank in our Bernoulli model. We use BIC criterion (see Equation (17)) to select the rank for our method. BTF\_Gradient does not provide rank selection, so we apply the same  $R$  outputted from our method to BTF\_Gradient. Because BTF\_Bayesian only supports the logistic link, we set the logistic links in all three BTF methods for fair comparison.

Figure 4a-b shows the performance comparison under the boolean model. We found that the three BTF methods outer-perform the BooleanTF in terms of RMSE. This demonstrates the advantage of probabilistic models for population inference. Because BTF methods are built upon a population model, they are able to estimate the population mean  $\mathbb{E}(\mathcal{Y})$  well. In contrast, the BooleanTF aims to find patterns in a specific data but does not address the population estimation. For MER, the BooleanTF performs reasonably well in distinguishing 0's vs. 1's (Figure 4 (b)). This confirms our earlier assertion on the data

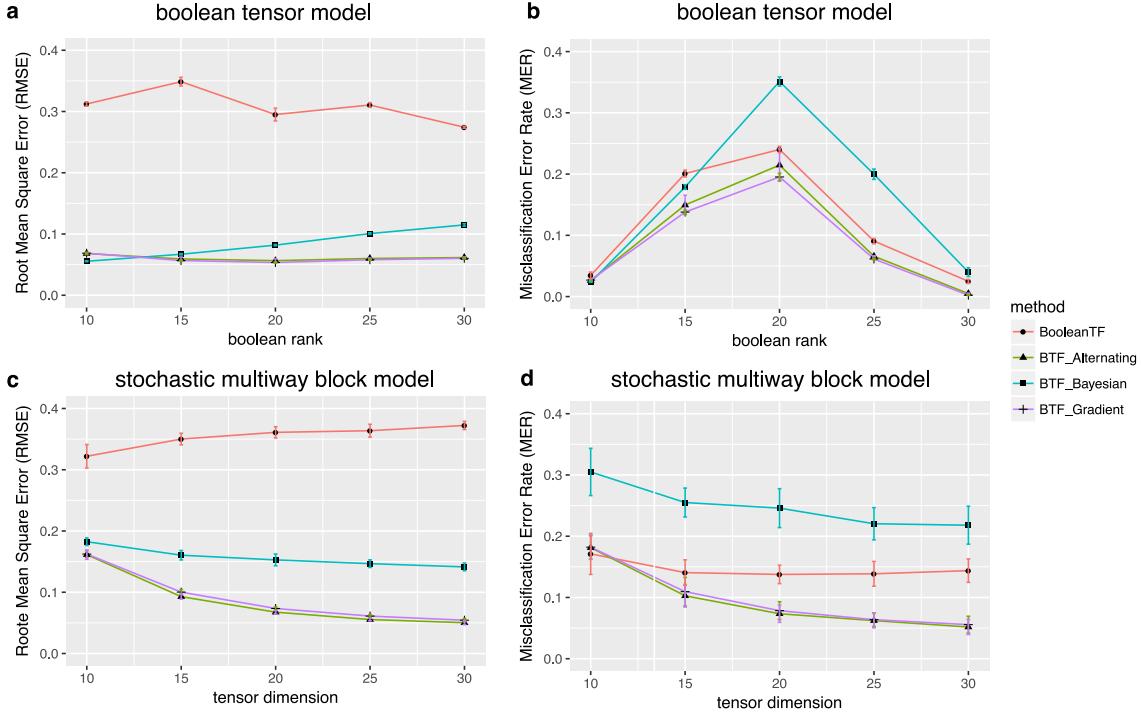


Figure 4: Performance comparison in simulations. Panels (a) and (b) show the estimation errors for tensors simulated from boolean tensor models, and Panels (c) and (d) are for stochastic multiway block models. The reported RMSE and MER are averaged over  $n_{\text{sim}} = 30$  replicates.

mining nature of the BooleanTF. Interesting, we found that the MER peaks at  $R = 20$ . Further investigation reveals that this setting corresponds to the case when the Bernoulli probabilities  $\mathbb{E}(\mathcal{Y})$  concentrate around 0.5. Since classification becomes harder when the Bernoulli probability is close to 0.5, we believe this is the reason driving the peak of MER. Indeed, the averaged Bernoulli probability for  $R = 10, 15, 20, 25, 30$  are 0.31, 0.44, 0.53, 0.61, 0.68, respectively. Figure 4b also shows that the two BTF methods with frequentist setup yield a lower classification error than the Bayesian. One possible reason is the normal prior in BTF\_Bayesian has a poor distinguishing power around  $\theta \approx 0$ , which corresponds to the hardest case when Bernoulli probability  $\approx 0.5 (= \text{logit}(\theta))$ .

### 5.2.2 STOCHASTIC MULTI-WAY BLOCK MODEL

We then simulate tensors from stochastic multi-way block model (SMBM). The SMBM can be viewed as a higher-order generalization of stochastic block model (SBM) commonly used for random graphs, network analysis, and community detection (Anandkumar et al., 2014;

(Abbe, 2017). A main feature of this simulated model is that the success probability remains the same within each block but differs from one block to another.

Specifically, we generate data tensors  $\mathcal{Y}$  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  $\Theta$ . The probability tensor  $\Theta$  is partitioned into five blocks along each of the modes,

$$\text{Probit}^{-1}(\Theta) = \mathcal{C} \times_1 \mathbf{N}_1 \times_2 \mathbf{N}_2 \times_3 \mathbf{N}_3 \quad (20)$$

where  $\mathcal{C} = [\![c_{m_1 m_2 m_3}]\!] \in \mathbb{R}^{5 \times 5 \times 5}$  is a core tensor corresponding to block-means in a probit scale,  $\mathbf{N}_1, \mathbf{N}_2, \mathbf{N}_3 \in \{0, 1\}^{5 \times d}$  are membership matrices indicating the block allocation along each of the mode, and  $m_1, m_2, m_3 = 1, \dots, 5$  are block indices. The block-means  $c_{m_1 m_2 m_3}$  are generated i.i.d. from  $\text{Unif}[-1, 1]$ . All other simulation settings are the same as in Section 5.2.1.

Figure 4c–d shows the error comparison under the SMBM. Again, we found that BooleanTF performs poorly in RMSE. This is probably due to the lack of population model in the BooleanTF. The two frequentist-type BTF methods exhibit decaying error with dimension. For the classification error, the booleanTF tends to perform well but still less favorably than the two BTF methods. The BTF\_Bayesian again exhibits a poor classification accuracy. This is possibly due to the same reason as we observed in the earlier simulations. The normal prior imposes a mass around the origin, making the classification harder near the boundary when probability  $\approx 0.5$ .

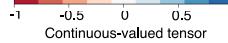
We would like to emphasize that we treat the BTF\_Gradient more as a complementary rather than a completing method. Since both BTF\_Gradient and BTF\_Alternating share the same models, they allow us to benchmark the statistical error. The good performance of both indicate that it is the statistical error that dominates the empirical accuracy. Our algorithm-independent results, such as statistical convergence rate, SNR phase diagram, and mini-max rate, equally apply to both methods. As such, our work can be viewed a unified framework for all type of this approach.

**Recovering latent tensors.** Lastly, we run an experiment to verify the capability of our method in recovering the latent tensor  $\Theta$  from binary observations. The set-up is the same as in the stochastic multi-way block models, except that we now modify the generation of block means  $\{c_{m_1 m_2 m_3}\}$  using the following three sub-models:

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

Table 3 reports the performance of our method, in terms of the relative loss, the estimated rank, and the running time, averaged over  $n_{\text{sim}} = 30$  data replications, for the three block tensor 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

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)



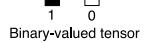


Table 3: Performance under the three-way block models. Columns 2–4 are color images of the simulated tensor under different block mean models. Reported are the relative loss, the estimated rank, and the running time, averaged over 30 data replications, with the standard error shown in the parenthesis.

illustration, we also plot one typical realization of the true signal tensor, the input binary tensor, and the recovered signal tensor for each model in Table 3. It is interesting to see that, not only the block structure but also the tensor magnitude are well recovered. Note 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 is generated from a probit model, but we always fit with a logistic link. Our method is shown to maintain a sound performance under potential model misspecification.

## 6. Real-world data

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 applied tensor decompositions 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 % test, 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 the root mean squared error (RMSE), averaged over five random splits of the training and testing data. It is clearly seen from this table that, the binary tensor decomposition substantially outperforms the classical continuous-valued tensor decomposition. In all datasets, the former obtains a substantially higher AUC and mostly a lower RMSE. We also report in Table 4 the percentage of 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 entries, 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 selected 5 clusters out of the 9 components. Figure 5a plots the nine tensor factors along the country mode. It

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

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. The evaluation criteria are averaged over five random splits of the training and testing data.

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 5b. 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 5b.

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 six 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 located 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.

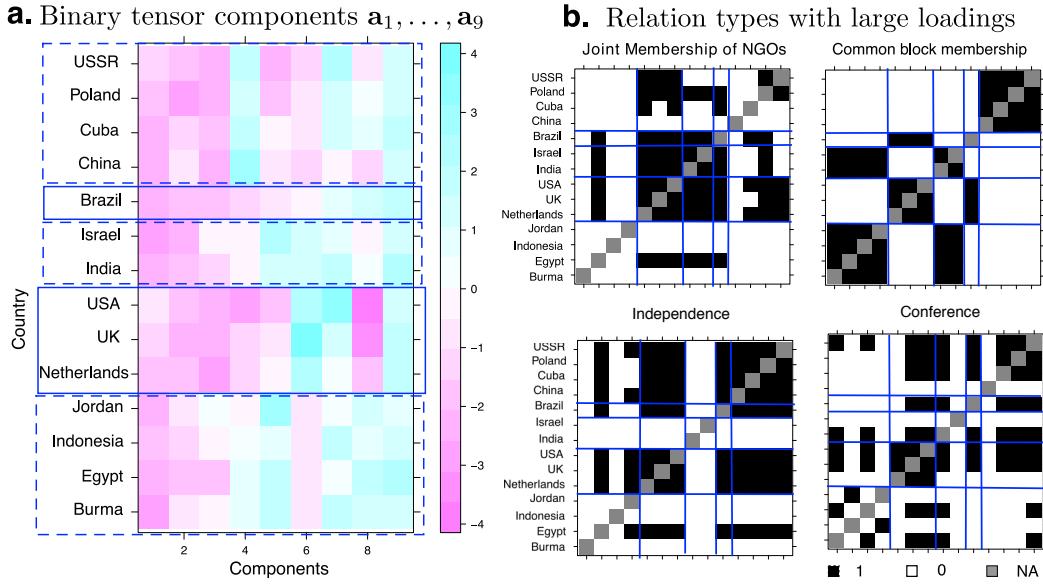


Figure 5: Analysis of the *Nations* dataset. (a) Binary tensor components. Top 9 tensor components in the country mode from the binary tensor decomposition. The overlaid box depicts the results from the  $K$ -means clustering. (b) Relation types with large loadings. Top four relationships identified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

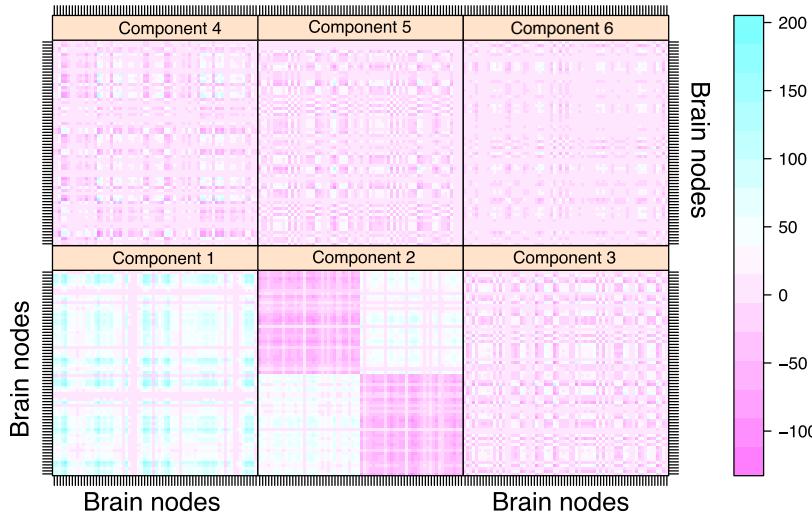


Figure 6: Analysis of the *HCP* data: heatmap for binary tensor components across brain regions. For each component, we plot the connection matrix  $\mathbf{A}_r = \lambda_r \mathbf{a}_r \otimes \mathbf{a}_r$  across the 68 brain nodes.

## 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 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, greater than or equal to  $\mathcal{L}_Y(\Theta_{\text{true}})$ . 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  $\Theta_{\text{true}}$ . 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.

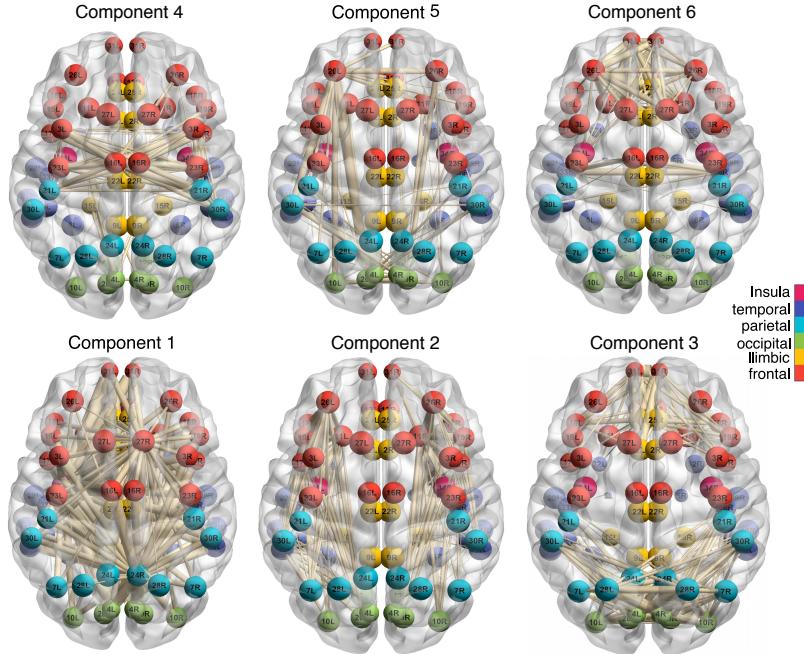


Figure 7: Analysis of the *HCP* data: edges with high loadings in the tensor decomposition components. We plot the top 10% edges with positive  $A_r(i, j)$ . The width of the edge is proportional to the magnitude of  $A_r(i, j)$ .

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 our 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.

## Appendices

### 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  $\mathcal{S} = \llbracket s_{i_1, \dots, i_K} \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$  is an order- $K$  tensor whose entries are independent random variables that satisfy*

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

*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  $\mathcal{S} = \llbracket s_{i_1, \dots, i_K} \rrbracket \in \mathbb{R}^{d_1 \times \dots \times d_K}$  is an order- $K$  tensor whose entries are independent random variables that satisfy*

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

*Then we have*

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

*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  $L^{-1}s_{i_1, \dots, i_K}$  is zero-mean and supported on  $[-1, 1]$ . Therefore,  $L^{-1}s_{i_1, \dots, i_K}$  is sub-Gaussian with parameter  $\frac{1-(-1)}{2} = 1$ , i.e.

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

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

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

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.  $\square$

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

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

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

*Proof.* Let  $\text{rank}(\cdot)$  denote the regular matrix rank, and  $\mathcal{A}_{(k)}$  denote the mode- $k$  matricization of  $\mathcal{A}$ ,  $k \in [K]$ . Define  $\text{rank}_T(\mathcal{A}) = (R_1, \dots, R_K)$  as the Tucker rank of  $\mathcal{A}$ , with  $R_k = \text{rank}(\mathcal{A}_{(k)})$ . The condition  $\text{rank}(\mathcal{A}) \leq R$  implies that  $R_k \leq R$  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

$$\|\mathcal{A}\|_* \leq \sqrt{\frac{\prod_{k=2}^K R_k}{\max_{k \geq 2} R_k}} \|\mathcal{A}_{(1)}\|_* \leq R^{\frac{K-2}{2}} \|\mathcal{A}_{(1)}\|_*, \quad (21)$$

where  $\mathcal{A}_{(1)}$  is a  $d_1$ -by- $\prod_{k \geq 2} d_k$  matrix with its rank bounded by  $R$ . Furthermore, the relationship between the matrix norms implies that  $\|\mathcal{A}_{(1)}\|_* \leq \sqrt{R} \|\mathcal{A}_{(1)}\|_F = \sqrt{R} \|\mathcal{A}\|_F$ . Combining this fact with the inequality (21) yields the final claim.  $\square$

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

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

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

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

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

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

which completes the proof.  $\square$

**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  $x > 0$ .  $\square$

**Lemma 6.** Let  $\mathcal{Y} \in \{0, 1\}^{d_1 \times \dots \times d_K}$  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  $\Theta$ . Let  $\mathbb{P}_0$  denote the distribution of  $\mathcal{Y}|0$  induced by the zero parameter tensor. Then

$$KL(\mathbb{P}_\Theta, \mathbb{P}_0) \leq 4\dot{f}^2(0)\|\Theta\|_F^2.$$

*Proof.* We have that

$$\begin{aligned} KL(\mathbb{P}_\Theta, \mathbb{P}_0) &= \sum_{i_1, \dots, i_K} KL(\mathcal{Y}_{i_1, \dots, i_K} | \theta_{i_1, \dots, i_K}, \mathcal{Y}_{i_1, \dots, i_K} | 0) \leq \sum_{i_1, \dots, i_K} \frac{(f(\theta_{i_1, \dots, i_K}) - f(0))^2}{f(0)(1-f(0))} \\ &= \sum_{i_1, \dots, i_K} \frac{\dot{f}^2(\eta_{i_1, \dots, i_K} \theta_{i_1, \dots, i_K}) (\theta_{i_1, \dots, i_K} - 0)^2}{f(0)(1-f(0))} \leq \sum_{i_1, \dots, i_K} 4\dot{f}^2(0)\theta_{i_1, \dots, i_K}^2 \\ &= 4\dot{f}^2(0)\|\Theta\|_F^2, \end{aligned}$$

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

**Lemma 7 (Varshamov-Gilbert bound).** Let  $\Omega = \{(w_1, \dots, w_m) : w_i \in \{0, 1\}\}$ . Suppose  $m > 8$ . Then there exists a subset  $\{w^{(0)}, \dots, w^{(M)}\}$  of  $\Omega$  such that  $w^{(0)} = (0, \dots, 0)$  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  $d_{\text{total}} = \prod_{k \geq 1} d_k$ . For any constant  $0 \leq \gamma \leq 1$ , there exist a finite set of tensors  $\mathcal{X} = \{\Theta_i : i = 1, \dots\} \subset \mathcal{D}(R, \alpha)$  satisfying the following four properties:

- (i)  $\text{Card}(\mathcal{X}) \geq 2^{Rd_1/8} + 1$ , where  $\text{Card}$  denotes the cardinality;
- (ii)  $\mathcal{X}$  contains the zero tensor  $\mathbf{0} \in \mathbb{R}^{d_1 \times \dots \times d_K}$ ;

- (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:

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

We then consider the associated set of block tensors:

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

where  $\Theta(\cdot, \cdot, i_3, \dots, i_K)$  denotes the matrix obtained from  $\Theta$  by varying the indices in the first two modes and fixing the indices in the other modes,  $\mathbf{O}$  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  $\Theta$  consists of blocks of the type  $\Theta(\cdot, \cdot, i_3, \dots, i_K)$  for any  $(i_3, \dots, i_K) \in [d_3] \times \dots \times [d_K]$ , stacked on top of each other, and furthermore, the block  $\Theta(\cdot, \cdot, i_3, \dots, i_K) \in \mathbb{R}^{d_1 \times d_2}$  is filled out by copying the matrix  $\mathbf{M} \in \mathbb{R}^{d_1 \times R}$  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  $\text{Card}(\mathcal{X}) \geq 2^{d_1 R/8} + 1$  containing the zero  $d_1 \times \dots \times d_K$  tensor, such that, for any two distinct elements  $\Theta_i$  and  $\Theta_j$  in  $\mathcal{X}$ ,

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

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

## B Proofs of main theorem

### B.1 PROOF OF THEOREM 1

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

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

Define

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

where the former is the collection of the score functions evaluated at  $\Theta_{\text{true}}$ , and the latter is the collection of the Hessian functions evaluated at  $\Theta_{\text{true}}$ . 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- $(d_1, \dots, d_K)$  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  $\Theta_{\text{true}}$  and obtain

$$\mathcal{L}_{\mathcal{Y}}(\Theta) = \mathcal{L}_{\mathcal{Y}}(\Theta_{\text{true}}) + \langle \mathcal{S}_{\mathcal{Y}}(\Theta_{\text{true}}), \Theta - \Theta_{\text{true}} \rangle + \frac{1}{2} \text{vec}(\Theta - \Theta_{\text{true}})^T \mathcal{H}_{\mathcal{Y}}(\check{\Theta}) \text{vec}(\Theta - \Theta_{\text{true}}), \quad (22)$$

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- $\prod_k d_k$  Hessian matrix evaluated at  $\check{\Theta}$ .

We first bound the linear term in (22). Note that, by Lemma 4,

$$|\langle \mathcal{S}_{\mathcal{Y}}(\Theta_{\text{true}}), \Theta - \Theta_{\text{true}} \rangle| \leq \|\mathcal{S}_{\mathcal{Y}}(\Theta_{\text{true}})\|_{\sigma} \|\Theta - \Theta_{\text{true}}\|_{*}. \quad (23)$$

Define

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

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

$$\mathbb{E}(s_{i_1, \dots, i_K}) = 0, \quad |s_{i_1, \dots, i_K}| \leq L_{\alpha}, \quad \text{for all } (i_1, \dots, i_K) \in [d_1] \times \dots \times [d_K]. \quad (24)$$

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

$$\|\mathcal{S}_{\mathcal{Y}}(\Theta_{\text{true}})\|_{\sigma} \leq C_2 L_{\alpha} \log K \sqrt{\sum_k d_k}, \quad (25)$$

where  $C_1, C_2$  are two positive constants. Furthermore, note that  $\text{rank}(\Theta) \leq R$ ,  $\text{rank}(\Theta_{\text{true}}) \leq R$ , so  $\text{rank}(\Theta - \Theta_{\text{true}}) \leq 2R$ . By lemma 3,  $\|\Theta - \Theta_{\text{true}}\|_* \leq (2R)^{\frac{K-1}{2}} \|\Theta - \Theta_{\text{true}}\|_F$ . Combining (23), (24) and (25), we have that, with probability at least  $1 - \exp(-C_1 \log K \sum_k d_k)$ ,

$$|\langle \mathcal{S}_{\mathcal{Y}}(\Theta_{\text{true}}), \Theta - \Theta_{\text{true}} \rangle| \leq C_2 L_{\alpha} \sqrt{R^{K-1} \sum_k d_k} \|\Theta - \Theta_{\text{true}}\|_F, \quad (26)$$

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

We next bound the quadratic term in (22). Let  $W = \text{vec}(\Theta - \Theta_{\text{true}})$ . Note that

$$\begin{aligned} \text{vec}(\Theta - \Theta_{\text{true}})^T \mathcal{H}_{\mathcal{Y}}(\check{\Theta}) \text{vec}(\Theta - \Theta_{\text{true}}) &= \sum_{i_1, \dots, i_K} \left( \frac{\partial^2 \mathcal{L}_{\mathcal{Y}}}{\partial \theta_{i_1, \dots, i_K}^2} \Big|_{\Theta=\check{\Theta}} \right) (\Theta_{i_1, \dots, i_K} - \Theta_{\text{true}, i_1, \dots, i_K})^2 \\ &\geq \gamma_{\alpha} \sum_{i_1, \dots, i_K} (\Theta_{i_1, \dots, i_K} - \Theta_{\text{true}, i_1, \dots, i_K})^2 \\ &= \gamma_{\alpha} \|\Theta - \Theta_{\text{true}}\|_F^2, \end{aligned} \quad (27)$$

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

Combining (22), (26) and (27), we have that, for all  $\Theta \in \mathcal{D}$ , with probability at least  $1 - \exp(-C_1 \log K \sum_k d_k)$ ,

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

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

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

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

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

Henceforth,

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

This completes the proof.  $\square$

**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  $\{\Theta \in \mathcal{D} : \mathcal{L}_Y(\Theta) \geq \mathcal{L}_Y(\Theta_{\text{true}})\}$ .

## B.2 PROOF OF THEOREM 2

*Proof.* Without loss of generality, we assume  $d_1 = d_{\text{max}}$ , 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  $\mathcal{X} = \{\Theta_i : i = 1, \dots\} \subset \mathcal{D}(R, \alpha)$  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),

$$\begin{aligned} \text{for the logistic link: } & \text{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \frac{4}{\sigma^2} \|\Theta\|_F^2, \\ \text{for the probit link: } & \text{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \frac{2}{\pi \sigma^2} \|\Theta\|_F^2, \\ \text{for the Laplacian link: } & \text{KL}(\mathbb{P}_\Theta, \mathbb{P}_{\mathbf{0}}) \leq \frac{1}{\sigma^2} \|\Theta\|_F^2, \end{aligned} \tag{28}$$

where  $\sigma$  is the scale parameter. Therefore, we have that the KL divergence between  $\mathbb{P}_\Theta$  and  $\mathbb{P}_0$  satisfies

$$\text{KL}(\mathbb{P}_\Theta, \mathbb{P}_0) \leq \frac{2}{\pi\sigma^2} \|\Theta\|_F^2 \leq \frac{1}{8\pi} R d_1 \gamma^2, \quad (29)$$

where the first inequality comes from (28), and the second inequality comes from property (iv) of  $\mathcal{X}$ . From (29) and the property (i), we deduce that the condition

$$\frac{1}{\text{Card}(\mathcal{X}) - 1} \sum_{\Theta \in \mathcal{X}} \text{KL}(\mathbb{P}_\Theta, \mathbb{P}_0) \leq \varepsilon \log \{\text{Card}(\mathcal{X}) - 1\} \quad (30)$$

is satisfied for any  $\varepsilon \geq 0$  if  $\gamma \in [0, 1]$  is chosen to be sufficiently small, depending on  $\varepsilon$ , e.g.,  $\gamma \leq \sqrt{3\varepsilon}$ . By applying Tsybakov (2009, Theorem 2.5) to (30), and in view of the property (iv), we obtain that

$$\inf_{\hat{\Theta}} \sup_{\Theta_{\text{true}} \in \mathcal{X}} \mathbb{P} \left( \|\hat{\Theta} - \Theta_{\text{true}}\|_F \geq \frac{\gamma}{8} \min \left\{ \alpha \sqrt{d_{\text{total}}}, \sigma \sqrt{R d_1} \right\} \right) \geq \frac{1}{2} \left( 1 - 2\varepsilon - \sqrt{\frac{16\varepsilon}{R d_1}} \right). \quad (31)$$

Note that  $\|\hat{\Theta} - \Theta_{\text{true}}\|_F = \sqrt{d_{\text{total}}} \text{Loss}(\hat{\Theta}, \Theta_{\text{true}})$  and  $\mathcal{X} \subset \mathcal{D}(R, \alpha)$ . Therefore, we conclude from (31) that

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

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  $\Theta \in \mathcal{X}$ ,  $\Theta$  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,  $\mathcal{Y} = [\![y_{i_1, \dots, i_K}]\!]|\Theta \sim_{\text{i.i.d.}} N(0, \sigma^2)$ . Note that, for the Gaussian distribution,

$$KL(\mathbb{P}_\Theta, \mathbb{P}_0) = \frac{\|\Theta\|_F}{2\sigma^2} \leq \frac{1}{16} R d_1 \gamma^2.$$

So the condition

$$\frac{1}{\text{Card}(\mathcal{X}) - 1} \sum_{\Theta \in \mathcal{X}} \text{KL}(\mathbb{P}_\Theta, \mathbb{P}_0) \leq \varepsilon \log (\text{Card}(\mathcal{X}) - 1) \quad (32)$$

is satisfied for any  $\varepsilon \geq 0$  if  $\gamma \in [0, 1]$  is chosen to be sufficiently small, depending on  $\varepsilon$ , e.g.,  $\gamma \leq \sqrt{3\varepsilon}$ . In view of the property (iv) and (32), 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 global convergence is similar to Zhou et al. (2013, Proposition 1). We present the main ideas here for completeness. By condition (A2), the block update is well-defined and differentiable. The isolation of stationary points ensures that there are only finite number of stationary points. We need to show that the every sub-sequence of  $\mathbf{A}^{(t)}$  converges to a same limiting point; i.e.,

$$\limsup_{t \rightarrow \infty} \mathbf{A}^{(t)} = \liminf_{t \rightarrow \infty} \mathbf{A}^{(t)}. \quad (33)$$

Let  $\mathbf{A}^{(t_n)}$  be a sub-sequence with limiting point  $\mathbf{A}^*$ . As the algorithm monotonically increases the objective value, the limiting point  $\mathbf{A}^*$  is a stationary point of  $\mathcal{L}$ . Now take the set of all limiting points. The set of limiting points are contained in the set  $\{\mathbf{A} : \mathcal{L}(\mathbf{A}) \geq \mathcal{L}(\mathbf{A}^{(0)})\}$  and thus also compact because of condition (A1). 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, (33) holds and  $\mathbf{A}^{(t)}$  converges to a stationary point of  $\mathcal{L}$ .

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

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

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

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

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

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

where  $C > 0$  is a constant. □

## B.5 PROOF OF THEOREM 4

*Proof.* Based on the Remark 2 for Theorem 1, we have

$$\text{Loss}(\Theta^*, \Theta_{\text{true}}) \leq \frac{C_2 L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}}. \quad (34)$$

Meanwhile, Proposition 1 implies that

$$\text{Loss}(\Theta^{(t)}, \Theta^*) \leq C_1 \rho^t \text{Loss}(\Theta^{(0)}, \Theta^*). \quad (35)$$

Combining (34) and (35) gives

$$\text{Loss}(\Theta^{(t)}, \Theta_{\text{true}}) \leq \text{Loss}(\Theta^{(t)}, \Theta^*) + \text{Loss}(\Theta_{\text{true}}, \Theta^*) \quad (36)$$

$$\leq C_1 \rho^t \text{Loss}(\Theta^{(0)}, \Theta^*) + \frac{C_2 L_\alpha}{\gamma_\alpha} \sqrt{\frac{R^{K-1} \sum_k d_k}{\prod_k d_k}}. \quad (37)$$

□

## 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 grant AG034570.

## References

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