Beyond the Signs: Nonparametric Tensor Completion via Sign Series

By Chanwoo Lee
Supporting document for the preliminary examination of
PhD degree in Statistics

Abstract

We consider the problem of tensor estimation from noisy observations with possibly missing entries. A nonparametric approach to tensor completion is developed based on a new model which we coin as sign representable tensors. The model represents the signal tensor of interest using a series of structured sign tensors. Unlike earlier methods, the sign series representation effectively addresses both low- and high-rank signals, while encompassing many existing tensor models—including CP models, Tucker models, single index models, structured tensors with repeating entries—as special cases. We provably reduce the tensor estimation problem to a series of structured classification tasks, and we develop a learning reduction machinery to empower existing low-rank tensor algorithms for more challenging high-rank estimation. Excess risk bounds, estimation errors, and sample complexities are established. We demonstrate the outperformance of our approach over previous methods on two datasets, one on human brain connectivity networks and the other on topic data mining.

1 Introduction

Higher-order tensors have recently received much attention in enormous fields including social networks (Anandkumar et al., 2014), neuroscience (Wang et al., 2017), and genomics (Hore et al., 2016). Tensor methods provide effective representation of the hidden structure in multiway data. In this paper we consider the signal plus noise model,

$$\mathcal{Y} = \Theta + \mathcal{E}, \tag{1}$$

where $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is an order-K data tensor, Θ is an unknown signal tensor of interest, and $\mathcal{E}$ is a noise tensor. Our goal is to accurately estimate Θ from the incomplete, noisy observation of $\mathcal{Y}$. In particular, we focus on the following two problems:

Q1 [Nonparametric tensor estimation]. How to flexibly estimate Θ under a wide range of structures, including both low-rankness and high-rankness?

Q2 [Complexity of tensor completion]. How many observed tensor entries do we need to consistently estimate the signal Θ?

Inadequacies of low-rank models. The signal plus noise model (2) is popular in tensor literature. Existing methods estimate the signal tensor based on low-rankness of Θ (Jain and Oh, 2014; Montanari and Sun, 2018). Common low-rank models include Canonical Polyadic (CP) tensors (Hitchcock, 1927), Tucker tensors (De Lathauwer et al., 2000), and block tensors (Wang and Zeng, 2019). While these methods have shown great success in theory, tensors in applications

often violate the low-rankness. Here we provide two examples to illustrate the limitation of classical models.

The first example reveals the sensitivity of tensor rank to order-preserving transformations. Let $\mathcal{Z} \in \mathbb{R}^{30\times 30\times 30}$ be an order-3 tensor with tensor rank$(\mathcal{Z}) = 3$ (formal definition is deferred to the end of this section). Suppose a monotonic transformation $f(z) = (1 + \exp(-cz))^{-1}$ is applied to $\mathcal{Z}$ entrywise, and we let the signal Θ in model (1) be the tensor after transformation. Figure 1a plots the numerical rank (see Section 7.1) of Θ versus c. As we see, the rank increases rapidly with c, rending traditional low-rank tensor methods ineffective in the presence of mild order-preserving nonlinearities. In digital processing (Ghadermarzy et al., 2018) and genomics analysis (Hore et al., 2016), the tensor of interest often undergoes unknown transformation prior to measurements. The sensitivity to transformation makes the low-rank model less desirable in practice.

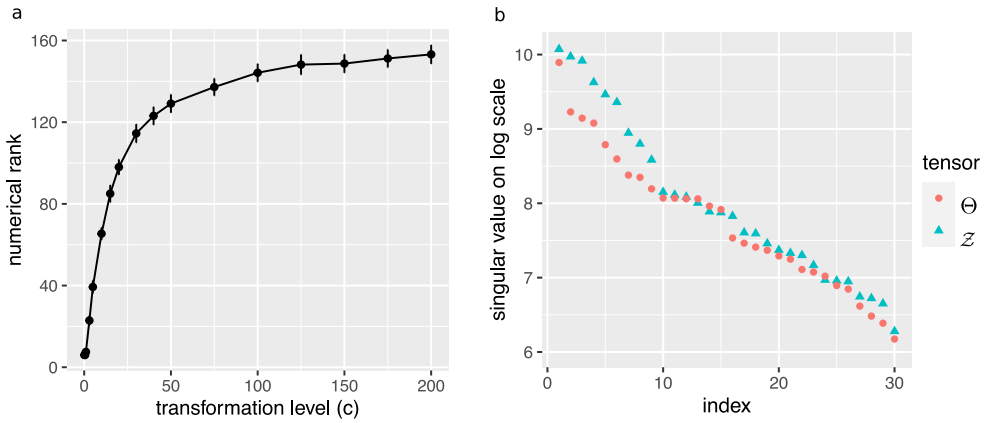


Figure 1: (a) Tensor rank vs. c in the first example. (b) Top $d = 30$ tensor singular values in the second example.

The second example demonstrates the inadequacy of classical low-rankness in representing special structures. Here we consider the signal tensor of the form $\Theta = \log(1 + \mathcal{Z})$, where $\mathcal{Z} \in \mathbb{R}^{d\times d\times d}$ is an order-3 tensor with entries $\mathcal{Z}(i, j, k) = d^{-1} \max(i, j, k)$ for $i, j, k \in \{1, \ldots, d\}$. The matrix analogy of Θ was studied in graphon analysis Chan and Airoldi (2014). In this case neither Θ nor $\mathcal{Z}$ is low-rank; in fact, the rank is no smaller than the dimension d as illustrated in Figure 1b. Again, classical low-rank models fail to address this type of tensor structure.

In the above and many other examples, the signal tensors Θ of interest have high rank. Classical low-rank models will miss these important structures. The observations have motivated us to develop more flexible tensor modeling.

Our contributions. We develop a new model called sign representable tensors to address the aforementioned challenges. Figure 2 illustrates our main idea. Our approach is built on the sign series representation of the signal tensor, and we propose to estimate the sign tensors through a series of weighted classifications. In contrast to existing methods, our method is guaranteed to recover a wide range of low- and high-rank signals. We highlight two main contributions that set our work apart from earlier literature.

Statistically, the problem of high-rank tensor estimation is challenging. Existing estimation theory (Anandkumar et al., 2014; Montanari and Sun, 2018; Cai et al., 2019) exclusively focuses on the regime of fixed r growing d. However, such premise fails in high-rank tensors, where the rank may grow with, or even exceed, the dimension. A proper notion of nonparametric complexity is crucial. We show that, somewhat surprisingly, the sign tensor series not only preserves all information in

the original signals, but also brings the benefits of flexibility and accuracy over classical low-rank models. The results fill the gap between parametric (low-rank) and nonparametric (high-rank) tensors, thereby greatly enriching the tensor model literature.

Computationally, a number of polynomial-time algorithms are readily available for 1-bit tensor estimation Wang and Li (2020); Han et al. (2020); Ghadermarzy et al. (2018). These algorithms enjoy computational efficiency while being restricted to binary inputs. Our work is orthogonal to these algorithm development, and we show that the high-rank tensor estimate is provably reducible to a series of binary tensor problems with carefully-designed weights. This reduction provides a generic engine to empower existing algorithms for a wider range of structured tensor problems. We use a divide-and-concur approach to combine efficient base algorithms, thereby achieving computational accuracy without the need to reinvent the wheel. The flexibility to import and adapt existing tensor algorithms is one advantage of our method.

We also highlight the challenges associated with tensors compared to matrices. High-rank matrix estimation is recently studied under nonlinear models (Ganti et al., 2015) and subspace clustering (Ongie et al., 2017; Fan and Udell, 2019). However, high-rank tensor problems is more challenging, because the tensor rank often exceeds the dimension when order K greater than two (Anandkumar et al., 2017). This is in sharp contrast to matrices ($K = 2$). We show that, applying matrix methods to higher-order tensors results in suboptimal estimates. A full exploitation of the higher-order structure is needed; this is another challenge we address in this paper.

Notation. We use $[n] = \{1, \dots, n\}$ for n-set with $n \in \mathbb{N}_+$, $a_n \lesssim b_n$ if $\lim_{n\to\infty} a_n/b_n \le c$ for some constant $c > 0$, and $a_n \asymp b_n$ if $c_1 \le \lim_{n\to\infty} a_n/b_n \le c_2$ for some constants $c_1, c_2 > 0$. We use $\mathcal{O}(\cdot)$ to denote the big-O notation, $\tilde{\mathcal{O}}(\cdot)$ the variant hiding logarithmic factors. Let $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ denote an order-K $(d_1, \dots, d_K)$-dimensional tensor, and $\Theta(\omega) \in \mathbb{R}$ denote the tensor entry indexed by $\omega \in [d_1] \times \cdots \times [d_K]$. An event A is said to occur "with very high probability" if $\mathbb{P}(A)$ tends to 1 faster than any polynomial of tensor dimension $d := \min_k d_k \to \infty$. The tensor rank (Hitchcock, 1927) is defined by $\mathrm{rank}(\Theta) = \min\{r \in \mathbb{N} \colon \Theta = \sum_{s=1}^{r} \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}\}$, where $\boldsymbol{a}_s^{(k)} \in \mathbb{R}^{d_k}$ are vectors for $k \in [K], s \in [r]$, and $\otimes$ denotes the outer product of vectors. We use $\mathrm{sgn}(\cdot) \colon \mathbb{R} \to \{-1, 1\}$ to denote the sign function, where $\mathrm{sgn}(y) = 1$ if $y \ge 0$ and -1 otherwise. We allow univariate functions, such as $\mathrm{sgn}(\cdot)$ and general $f \colon \mathbb{R} \to \mathbb{R}$, to be applied to tensors in an element-wise manner.

2 Model and proposal overview

Let $\mathcal{Y}$ be an order-K $(d_1, \dots, d_K)$-dimensional tensor generated from the model

$$\mathcal{Y} = \Theta + \mathcal{E}, \tag{2}$$

where $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ is an unknown signal tensor, and $\mathcal{E}$ is a noise tensor consisting of zero-mean, independent but not necessarily identically distributed entries. We allow heterogenous noise, in that the marginal distribution of noise entry $\mathcal{E}(\omega)$ may depend on ω. For a cleaner exposition, we assume the noise is bounded and the range of Y is in $[-1, 1]$; the extension to a sub-Gaussian noise is provided in Section 7.4. Our observation is an incomplete data tensor from (2), denoted $\mathcal{Y}_\Omega$, where $\Omega \subset [d_1] \times \cdots \times [d_K]$ is the index set of observed entries. We consider a general model on Ω that allows both uniform and non-uniform samplings. Specifically, let $\Pi = \{p_\omega\}$ be an arbitrarily predefined probability distribution over the full index set with $\sum_{\omega\in[d_1]\times\cdots\times[d_K]} p_\omega = 1$. We use $\omega \sim \Pi$ to denote the sampling rule, meaning ω in Ω are i.i.d. draws with replacement from distribution Π. The goal is to estimate Θ from Y_Ω. Note that Θ is not necessarily low-rank.

Proposal intuition. Before describing our main results, we provide the intuition behind our method. In the two examples in Section 1, the high-rankness in the signal Θ makes the estimation challenging. Now let us examine the sign of the π-shifted signal $\mathrm{sgn}(\Theta - \pi)$ for any given $\pi \in [-1, 1]$. It turns out that, these sign tensors share the same sign patterns as low-rank tensors. Indeed, the signal tensor in the first example has the same sign pattern as a rank-4 tensor, since $\mathrm{sgn}(\Theta - \pi) = \mathrm{sgn}(\mathcal{Z} - f^{-1}(\pi))$. The signal tensor in the second example has the same sign pattern as a rank-2 tensor, since $\mathrm{sgn}(\Theta(i,j,k) - \pi) = \mathrm{sgn}(\max(i,j,k) - d(e^{\pi} - 1))$ (see Example 5 in Section 3).

The above observation suggests a general framework to estimate both low- and high-rank signal tensors. Figure 2 illustrates the main crux of our method. We propose to estimate the signal tensor Θ by taking the average over structured sign tensors

$$\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}(\hat{\mathcal{Z}}_\pi), \text{ where } \hat{\mathcal{Z}}_\pi = \underset{\text{low rank tensor } \mathcal{Z}}{\arg\min} \text{ Weighted-Loss}(\mathrm{sgn}(\mathcal{Z}), \mathrm{sgn}(\mathcal{Y}_\Omega - \pi)). \quad (3)$$

Here $\mathrm{sgn}(\hat{\mathcal{Z}}_\pi) \in \{-1, 1\}^{d_1 \times \cdots \times d_K}$ is the sign tensor estimated at a series of $\pi \in \mathcal{H} = \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$, and Weighted-Loss$(\cdot, \cdot)$ denotes a classification objective function with an entry-specific weight to each tensor entry; its specific form will be described in Section 3.2. To obtain $\mathrm{sgn}(\hat{\mathcal{Z}}_\pi)$ for a given π, we propose to dichotomize the data tensor into a sign tensor $\mathrm{sgn}(\mathcal{Y}_\Omega - \pi)$ and estimate the de-noised sign by performing weighted classification.

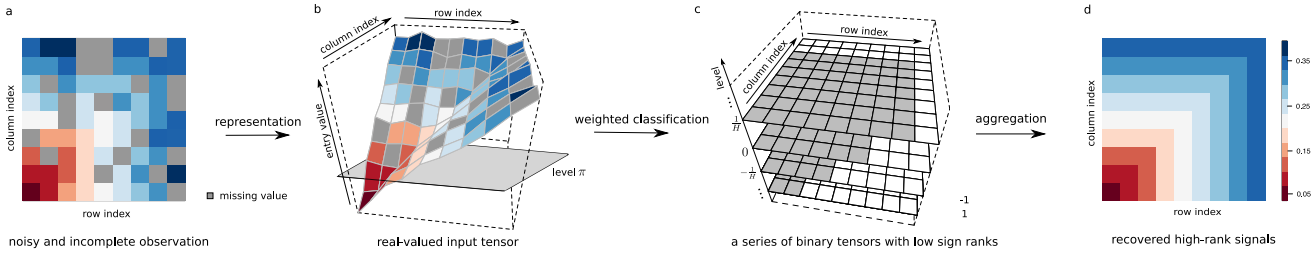


Figure 2: Illustration of our method in the context of an order-2 tensor (a.k.a. matrix). (a): a noisy, incomplete tensor input. (b)-(c): Estimation of sign tensor series $\mathrm{sgn}(\Theta - \pi)$ for $\pi \in \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$. (d): recovered signal $\hat{\Theta}$. The depicted signal is a full-rank matrix based on Example 5 in Section 3.

Our approach is built on the nonparametric sign representation of signal tensors. We show that a careful aggregation of dichotomized data not only preserves all information in the original signals, but also brings benefits of accuracy and flexibility over classical low-rank models. Unlike traditional methods, the sign representation is guaranteed to recover both low- and high-rank signals. In addition, a total of $H = \mathrm{poly}(d)$ dichotomized problems suffice to recover Θ under the considered model. The method therefore enjoys both statistical effectiveness and computational efficiency.

3 Oracle properties of sign representable tensors

This section develops sign representable tensor models for Θ in (2). We characterize the algebraic and statistical properties of sign tensor series, which serves the foundation for our method.

3.1 Sign-rank and sign tensor series

Let Θ be the tensor of interest, and $\mathrm{sgn}(\Theta)$ the corresponding sign pattern. The sign patterns induce an equivalence relationship between tensors. Two tensors are called sign equivalent, denoted $\simeq$, if they have the same sign pattern.

Definition 1 (Sign-rank)**.** The sign-rank of a tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is defined by the minimal rank among all tensors that share the same sign pattern as Θ; i.e.,

$$\operatorname{srank}(\Theta) = \min\{\operatorname{rank}(\Theta') \colon \Theta' \simeq \Theta,\ \Theta' \in \mathbb{R}^{d_1 \times \cdots \times d_K}\}.$$

This concept is important in combinatorics (Cohn and Umans, 2013), complexity theory (Alon et al., 2016), and quantum mechanics (De Wolf, 2003); we extend the notion to continuous-valued tensors. Note that the sign-rank concerns only the sign pattern but discards the magnitude information of Θ. In particular, $\operatorname{srank}(\Theta) = \operatorname{srank}(\operatorname{sgn}\Theta)$.

Like most tensor problems (Hillar and Lim, 2013), determining the sign-rank is NP hard in the worst case (Alon et al., 2016). Fortunately, tensors arisen in applications often possess special structures that facilitate analysis. The sign-rank is upper bounded by tensor rank. More generally, we show the following properties.

Proposition 1 (Upper bounds of the sign-rank)**.**
(a) [Upper bounds] For any strictly monotonic function $g \colon \mathbb{R} \to \mathbb{R}$ with $g(0) = 0$, we have $\operatorname{srank}(\Theta) \le \operatorname{rank}(g(\Theta))$.
(b) [Broadness] For every order $K \ge 2$ and dimension d, there exist tensors $\Theta \in \mathbb{R}^{d \times \cdots \times d}$ such that $\operatorname{rank}(\Theta) \ge d$ *but* $\operatorname{srank}(\Theta - \pi) \le 2$ *for all $\pi \in \mathbb{R}$.*

Propositions 1 demonstrates the strict broadness of low sign-rank family over the usual low-rank family. In particular, the sign-rank can be much smaller than the tensor rank, as we have shown in the two examples of Section 1. We provide several additional examples in Section 7.2 in which the tensor rank grows with dimension d but the sign-rank remains a constant. The results highlight the advantages of using sign-rank in the high-dimensional tensor analysis.

We now introduce a tensor family, which we coin as "sign representable tensors".

Definition 2 (Sign representable tensors)**.** Fix a level $\pi \in [-1, 1]$. A tensor Θ is called (r, π)-sign representable, if the tensor $(\Theta - \pi)$ has sign-rank bounded by r. A tensor Θ is called r-sign (globally) representable, if Θ is (r, π)-sign representable for all $\pi \in [-1, 1]$. The collection $\{\operatorname{sgn}(\Theta - \pi) \colon \pi \in [-1, 1]\}$ is called the sign tensor series. We use $\mathscr{P}_{\operatorname{sgn}}(r) = \{\Theta \colon \max_{\pi \in [-1,1]} \operatorname{srank}(\Theta - \pi) \le r\}$ to denote the r-sign representable tensor family.

We next show that the r-sign representable tensor family is a general model that incorporates most existing tensor models, including low-rank tensors, single index models, GLM models, and structured tensors with repeating entries.

Example 1 (CP/Tucker low-rank models)**.** The CP and Tucker low-rank tensors are the two most popular tensor models (Kolda and Bader, 2009). Let Θ be a low-rank tensor with CP rank r. We see that Θ belongs to the sign representable family; i.e., $\Theta \in \mathscr{P}_{\operatorname{sgn}}(r + 1)$ (the constant 1 is due to $\operatorname{rank}(\Theta - \pi) \le r + 1$). Similar results hold for Tucker low-rank tensors $\Theta \in \mathscr{P}_{\operatorname{sgn}}(r + 1)$, where $r = \prod_k r_k$ with r_k being the k-th mode Tucker rank of Θ.

Example 2 (Tensor block models (TBMs))**.** Tensor block model (Wang and Zeng, 2019; Chi et al., 2020) assumes a checkerboard structure among tensor entries under marginal index permutation. The signal tensor Θ takes at most r distinct values, where r is the total number of multiway blocks. Our model incorporates TBM because $\Theta \in \mathscr{P}_{\operatorname{sgn}}(r)$.

Example 3 (Generalized linear models (GLMs))**.** Let $\mathcal{Y}$ be a binary tensor from a logistic model (Wang and Li, 2020) with mean $\Theta = \operatorname{logit}(\mathcal{Z})$, where $\mathcal{Z}$ is a latent low-rank tensor. Notice that

Θ itself may be high-rank (see Figure 1a). By definition, Θ is a low-rank sign representable tensor. Same conclusion holds for general exponential-family models with a (known) link function (Hong et al., 2020).

Example 4 (Single index models (SIMs))**.** Single index model is a flexible semiparametric model proposed in economics (Robinson, 1988) and high-dimensional statistics (Balabdaoui et al., 2019; Ganti et al., 2017). The SIM assumes the existence of a (unknown) monotonic function $g\colon \mathbb{R} \to \mathbb{R}$ such that $g(\Theta)$ has rank r. We see that Θ belongs to the sign representable family; i.e., $\Theta \in \mathscr{P}_{\text{sgn}}(r+1)$.

Example 5 (Structured tensors with repeating entries)**.** Here we revisit the model introduced in Figure 1b of Section 1. Let Θ be an order-K tensor with entries $\Theta(i_1, \ldots, i_K) = \log(1 + \max_k x_{i_k}^{(k)})$, where $x_{i_k}^{(k)}$ are given numbers in $[0, 1]$ for all $i_k \in [d_k], k \in [K]$. We conclude that $\Theta \in \mathscr{P}_{\text{sgn}}(2)$, because the sign tensor $\text{sgn}(\Theta - \pi)$ with an arbitrary $\pi \in (0,\ \log 2)$ is a block tensor with at most two blocks (see Figure 2c). Similar results extend to structured tensors with entries $\Theta(i_1, \ldots, i_K) = g(\max_k x_{i_k}^{(k)})$, where $g(\cdot)$ is a polynomial of degree r. In this case, Θ is a high-rank tensor with at most $d_{\max}$ distinct entries but we have $\Theta \in \mathscr{P}_{\text{sgn}}(2r)$ (see proofs in Section 7.2).

3.2 Statistical characterization of sign tensors via weighted classification

We now provide the explicit form of the weighted loss introduced in (3), and show that sign tensors are characterized by weighted classification. The results bridge the algebraic and statistical properties of sign representable tensors.

For a given $\pi \in [-1, 1]$, define a π-shifted data tensor $\bar{\mathcal{Y}}_\Omega$ with entries $\bar{\mathcal{Y}}(\omega) = (\mathcal{Y}(\omega) - \pi)$ for $\omega \in \Omega$. We propose a weighted classification objective function

$$L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \underbrace{|\bar{\mathcal{Y}}(\omega)|}_{\text{weight}} \times \underbrace{|\text{sgn}\mathcal{Z}(\omega) - \text{sgn}\bar{\mathcal{Y}}(\omega)|}_{\text{classification loss}}, \tag{4}$$

where $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the decision variable to be optimized, $|\bar{\mathcal{Y}}(\omega)|$ is the entry-specific weight equal to the distance from the tensor entry to the target level π. The entry-specific weights incorporate the magnitude information into classification, where entries far away from the target level are penalized more heavily in the objective. In the special case of binary tensor $\mathcal{Y} \in \{-1, 1\}^{d_1 \times \cdots \times d_K}$ and target level $\pi = 0$, the loss (4) reduces to usual classification loss.

Our proposed weighted classification function (4) is important for characterizing $\text{sgn}(\Theta - \pi)$. Define the weighted classification risk

$$\text{Risk}(\mathcal{Z}) = \mathbb{E}_{\mathcal{Y}_\Omega} L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega), \tag{5}$$

where the expectation is taken with respect to $\mathcal{Y}_\Omega$ under model (2) and the sampling distribution $\omega \sim \Pi$. The form of $\text{Risk}(\cdot)$ implicitly depends on π; we suppress π when no confusion arises.

Proposition 2 (Global optimum of weighted risk)**.** *Suppose the data $\mathcal{Y}_\Omega$ is generated from model* (2) *with $\Theta \in \mathscr{P}_{\text{sgn}}(r)$. Then, for all $\bar{\Theta}$ that are sign equivalent to* $\text{sgn}(\Theta - \pi)$*,*

$$\text{Risk}(\bar{\Theta}) = \inf\{\text{Risk}(\mathcal{Z})\colon \mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}\} = \inf\{\text{Risk}(\mathcal{Z})\colon \text{rank}(\mathcal{Z}) \le r\}.$$

The results show that the sign tensor $\text{sgn}(\Theta - \pi)$ optimizes the weighted classification risk. This fact suggests a practical procedure to estimate $\text{sgn}(\Theta - \pi)$ via empirical risk optimization of $L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega)$. In order to establish the recovery guarantee, we shall address the uniqueness (up to sign equivalence) for the optimizer of $\text{Risk}(\cdot)$. The local behavior of Θ around π plays a key role in the accuracy.

Some additional notation is needed for stating the results in full generality. Let $d_{\text{total}} = \prod_{k=1}^{K} d_k$ denote the total number of tensor entries, and $\Delta s = 1/d_{\text{total}}$ a small tolerance. We quantify the distribution of tensor entries $\Theta(\omega)$ using a pseudo density, i.e., histogram with bin width $2\Delta s$. Let $G(\pi) := \mathbb{P}_{\omega\sim\Pi}[\Theta(\omega) \le \pi]$ denote the cumulative distribution function (CDF) of $\Theta(\omega)$ under $\omega \sim \Pi$. We partition $[-1, 1] = \mathcal{N}^c \cup \mathcal{N}$, such that the pseudo density based on 2Δ-bin is uniformly bounded over $\mathcal{N}^c$; i.e,

$$\mathcal{N}^c = \left\{ \pi \in [-1, 1] \colon \frac{G(\pi + \Delta s) - G(\pi - \Delta s)}{\Delta s} \le C \right\}, \text{ for some universal constant } C > 0,$$

and $\mathcal{N}$ otherwise. Both Θ and its induced CDF G implicitly depend on the tensor dimension.

Assumption 1 (α-smoothness)**.** Fix $\pi \in \mathcal{N}^c$. Assume there exist constants $\alpha = \alpha(\pi) > 0, c = c(\pi) > 0$, independent of tensor dimension, such that,

$$\sup_{\Delta s \le t < \rho(\pi, \mathcal{N})} \frac{G(\pi + t) - G(\pi - t)}{t^\alpha} \le c, \tag{6}$$

where $\rho(\pi, \mathcal{N}) := \min_{\pi' \in \mathcal{N}} |\pi - \pi'| + \Delta s$ denotes the adjusted distance from π to the nearest point in $\mathcal{N}$. The largest possible $\alpha = \alpha(\pi)$ in (6) is called the smoothness index at level π. We make the convention that $\alpha = \infty$ if the numerator in (6) is zero. A tensor Θ is called α-globally smooth, if (6) holds with global constants $\alpha > 0, c > 0$ for all $\pi \in \mathcal{N}^c$.

The smoothness index α quantifies the intrinsic hardness of recovering $\operatorname{sgn}(\Theta - \pi)$ from $\text{Risk}(\cdot)$. The value of α depends on both the sampling distribution $\omega \sim \Pi$ and the behavior of $\Theta(\omega)$. The recovery is easier at levels where points are less concentrated around π with a large value of $\alpha > 1$, or equivalently, when $G(\pi)$ remains almost flat around π. A small value of $\alpha < 1$ indicates the nonexistent (infinite) density at level π, or equivalently, when the $G(\pi)$ jumps by greater than the tolerance Δs at π. Table 1 illustrates the $G(\pi)$ for various models of Θ (see Section 5).

We now reach the main theorem in this section. For two tensors Θ_1, Θ_2, define the mean absolute error (MAE) as $\text{MAE}(\Theta_1, \Theta_2) = \mathbb{E}_{\omega\sim\Pi}|\Theta_1(\omega) - \Theta_2(\omega)|$.

Theorem 1 (Identifiability)**.** *Assume $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ is α-globally smooth. Then, for all $\pi \in \mathcal{N}^c$ and tensors $\bar{\Theta} \simeq \operatorname{sgn}(\Theta - \pi)$, we have*

$$\text{MAE}(\operatorname{sgn}\mathcal{Z}, \operatorname{sgn}\bar{\Theta}) \lesssim C(\pi) \left[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})\right]^{\alpha/(\alpha+1)} + \Delta s, \quad \textit{for all } \mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K},$$

where $C(\pi) > 0$ is independent of $\mathcal{Z}$.

The result establishes the recovery stability of sign tensors $\operatorname{sgn}(\Theta - \pi)$ using optimization with population risk (5). The bound immediately shows the uniqueness of the optimizer for $\text{Risk}(\cdot)$ up to a Δs-measure set under Π. We find that a higher value of α implies more stable recovery, as intuition would suggest. Similar results hold for optimization with sample risk (4) (see Section 4).

We conclude this section by applying Assumption 1 to the examples described in Section 3.1. For simplicity, suppose Π is the uniform sampling. The tensor block model is ∞-globally smooth. This is because the set $\mathcal{N}$ consists of finite $2\Delta s$-bin's covering the distinct block means in Θ. Furthermore, we have $\alpha = \infty$ for all $\pi \in \mathcal{N}^c$, since the numerator in (6) is zero. Similarly, the high-rank (d, d, d)-dimensional tensor $\Theta(i, j, k) = \log(1 + \frac{1}{d}\max(i, j, k))$ is ∞-globally smooth because $\alpha = \infty$ for all π except those in $\mathcal{N}$, where $\mathcal{N}$ collects d many $2\Delta s$-bin's covering $\log(1 + i/d)$ for all $i \in [d]$.

4 Nonparametric tensor completion via sign series

In previous sections we have established the sign series representation and its relationship to classification. In this section, we present our learning reduction proposal in details (Figure 2). We provide the estimation error bound and address the empirical implementation of the method.

4.1 Statistical error and sample complexity

Given a noisy incomplete tensor observation $\mathcal{Y}_\Omega$ from model (2), we cast the problem of estimating Θ into a series of weighted classifications. Specifically, we propose the signal tensor estimate using averaged structured sign tensors

$$\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \operatorname{sgn} \hat{\mathcal{Z}}_\pi, \quad \text{with} \quad \hat{\mathcal{Z}}_\pi = \underset{\mathcal{Z}:\, \operatorname{rank} \mathcal{Z} \le r}{\arg\min} L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi), \tag{7}$$

where $\mathcal{H} = \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$ is the series of levels to aggregate, $L(\cdot, \cdot)$ denotes the weighted classification objective defined in (4), and the rank constraint on $\mathcal{Z}$ follows from Proposition 2. For the theory, we assume the true r is known; in practice, r could be chosen in a data adaptive fashion via cross-validation or elbow method (Hastie et al., 2009).

The next theorem establishes the statistical convergence for the sign tensor estimate (7).

Theorem 2 (Sign tensor estimation)**.** *Suppose $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$ and $\Theta(\omega)$ is α-globally smooth under $\omega \sim \Pi$. Let $\hat{\mathcal{Z}}_\pi$ be the estimate in* (7)*, $d = \max_{k \in [K]} d_k$, and $t_d = \frac{dr \log |\Omega|}{|\Omega|} \lesssim 1$. Then, for all $\pi \in \mathcal{N}^c$, with very high probability over $\mathcal{Y}_\Omega$,*

$$\mathrm{MAE}(\operatorname{sgn} \hat{\mathcal{Z}}_\pi, \operatorname{sgn}(\Theta - \pi)) \lesssim t_d^{\alpha/(\alpha+2)} + \frac{1}{\rho^2(\pi, \mathcal{N})} t_d. \tag{8}$$

Theorem 2 provides the error bound for the sign tensor estimation. Compared to the population results in Theorem 1, we explicitly reveal the dependence of accuracy on the sample complexity and the level π. The result demonstrates the polynomial decay of sign errors with $|\Omega|$. Our sign estimate achieves consistent recovery using as few as $\tilde{\mathcal{O}}(dr)$ noisy entries.

Recall that $\mathcal{N}$ collects the levels for which the sign tensor is possibly nonrecoverable. Let $|\mathcal{N}|$ be the covering number of $\mathcal{N}$ with $2\Delta s$-bin's, i.e, $|\mathcal{N}| = \lceil \mathrm{Leb}(\mathcal{N})/2\Delta s \rceil$, where $\mathrm{Leb}(\cdot)$ is the Lebesgue measure and $\lceil \cdot \rceil$ is the ceiling function. Combining the sign representability of the signal tensor and the sign estimation accuracy, we obtain our main results on nonparametric tensor estimation.

Theorem 3 (Tensor estimation error)**.** *Consider the same conditions of Theorem 2. Let $\hat{\Theta}$ be the estimate in* (7)*. For any resolution parameter $H \in \mathbb{N}_+$, with very high probability over $\mathcal{Y}_\Omega$,*

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim (t_d \log H)^{\frac{\alpha}{\alpha+2}} + \frac{1 + |\mathcal{N}|}{H} + t_d H \log H. \tag{9}$$

In particular, setting $H = (1 + |\mathcal{N}|)^{1/2} t_d^{-1/2} \asymp \mathit{poly}(d)$ yields the tightest upper bound in (52)*.*

Theorem 3 demonstrates the convergence rate of our tensor estimation. The bound (52) reveals three sources of errors: the estimation error for sign tensors, the bias from sign series representations, and the variance thereof. The resolution parameter H controls the bias-variance tradeoff. We remark that the signal estimation error (52) is generally no better than the corresponding sign error (8). This is to be expected, since magnitude estimation is a harder problem than sign estimation.

In the special case of full observation with equal dimension $d_1 = \cdots = d_K = d$ and bounded $|\mathcal{N}| \leq C$, our signal estimate achieves convergence

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim r d^{-(K-1)\min(\frac{\alpha}{\alpha+2}, \frac{1}{2})} \log^2 d,$$

by setting $H \asymp d^{(K-1)/2}$. Compared to earlier methods, our estimation accuracy applies to both low- and high-rank signal tensors. The rate depends on the sign complexity $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$, and this r is often much smaller than the usual tensor rank (see Section 3.1). Our result also reveals that the convergence becomes favorable as the order of data tensor increases.

We apply our method to the main examples in Section 3.1, and compare the results with existing literature. The numerical comparison is provided in Section 5.

Example 2 (TBMs)**.** Consider a tensor block model with r multiway blocks. Our result implies a rate $\tilde{\mathcal{O}}(d^{-(K-1)/2})$ by taking $\alpha = \infty$ and $|\mathcal{N}| \leq r^K \lesssim \mathcal{O}(1)$. This rate agrees with the previous root-mean-square error (RMSE) for block tensor estimation (Wang and Zeng, 2019).

Example 3 (GLMs)**.** Consider a GLM tensor $\Theta = g(\mathcal{Z})$, where g is a known link function and $\mathcal{Z}$ is a latent low-rank tensor. Suppose the CDF of $\Theta(\omega)$ is uniformly bounded as $d \to \infty$. Applying our results with $\alpha = 1$ and finite $|\mathcal{N}|$ yields $\tilde{\mathcal{O}}(d^{-(K-1)/3})$. This rate is slightly slower than the parametric RMSE rate (Zhang and Xia, 2018; Wang and Li, 2020), as expected. The reason is that our estimate remains valid for unknown g and general high-rank tensors. The nonparametric rate is the price one has to pay for not knowing the form $\Theta = g(\mathcal{Z})$ as a priori.

Example 4 (SIMs)**.** The earlier example has shown the nonparametric rate $\tilde{\mathcal{O}}(d^{-(K-1)/3})$ when applying our method to single index tensor model. In the matrix case with $K = 2$, our result yields error rate $\tilde{\mathcal{O}}(d^{-1/3})$, which is faster than the RMSE rate $\mathcal{O}(d^{-1/4})$ obtained by Ganti et al. (2015).

Example 5 (Structured tensors with repeating entries)**.** We consider a more general model than that in Section 1. Consider a r-sign representable tensor $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$ with at most d distinct entries with repetition pattern. Applying our results with $\alpha = \infty$ and $|\mathcal{N}| = d$ yields the rate $\tilde{\mathcal{O}}(d^{-(K-2)/2})$.

The following corollary reveal the sample complexity for nonparamtric tensor completion.

Corollary 1 (Sample complexity for nonparametric completion)**.** *Assume the same conditions of Theorem 3 and bounded $|\mathcal{N}|$. Then, with high probability over $\mathcal{Y}_\Omega$,*

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \to 0, \quad \textit{as} \quad \frac{|\Omega|}{dr \log^2 |\Omega|} \to \infty.$$

Our result improves earlier work (Yuan and Zhang, 2016; Ghadermarzy et al., 2019; Lee and Wang, 2020) by allowing both low- and high-rank signals. Interestingly, the sample requirements depend only on the sign complexity dr but not the nonparametric complexity α. Note that $\tilde{\mathcal{O}}(dr)$ roughly matches the degree of freedom of sign tensors, suggesting the optimality of our sample requirements.

4.2 Implementation via learning reduction

This section addresses the practical implementation of our estimation (7). We take a learning reduction approach by dividing the full procedure into a meta algorithm and $2H+1$ base algorithms. The meta algorithm takes the average of $(2H+1) \asymp \mathrm{poly}(d)$ sign tensors, whereas each base

Algorithm 1 Nonparametric tensor completion via learning reduction

Input: Noisy and incomplete data tensor $\mathcal{Y}_\Omega$, rank r, resolution parameter H, ridge penalty λ.

1: **for** $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$ **do**

2: Define π-shifted tensor $\bar{\mathcal{Y}} = \mathcal{Y} - \pi$ and corresponding sign tensor $\mathrm{sgn}(\bar{\mathcal{Y}}) = \mathrm{sgn}(\mathcal{Y} - \pi)$.

3: Perform 1-bit tensor estimation algorithm (Ghadermarzy et al., 2018; Wang and Li, 2020; Hong et al., 2020; Alquier et al., 2019) on $\bar{\mathcal{Y}}_\Omega$ and obtain

$$\hat{\mathcal{Z}}_\pi \leftarrow \underset{\text{low-rank } \mathcal{Z}}{\arg\min} \sum_{\omega \in \Omega} |\bar{\mathcal{Y}}(\omega)| F(\mathcal{Z}(\omega) \mathrm{sgn} \bar{\mathcal{Y}}(\omega)) + \lambda \|\mathcal{Z}\|_F^2,$$

where $F(\cdot)$ is the large-margin loss and λ is the penalty parameter.

4: **end for**

Output: Estimated signal tensor $\hat{\Theta}_F = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}(\hat{\mathcal{Z}}_\pi)$.

algorithm estimates the tensor $\mathrm{sgn}(\Theta - \pi)$ given binary input $\mathrm{sgn}(\mathcal{Y} - \pi)$ and a target rank r. The full procedure is described in Algorithm 1 and Figure 2.

The base algorithm reduces to a low-rank 1-bit tensor estimation problem. Following the common practice in classification Bartlett et al. (2006), we replace the 0-1 loss $\ell(z, y) = |\mathrm{sgn} z - \mathrm{sgn} y|$ in (4) with a continuous large-margin loss $F(m)$ where $m = z\mathrm{sgn}(y)$ is the margin. Examples of large-margin loss are hinge loss $F(m) = (1 - m)_+$, logistic loss $F(m) = \log(1 + e^{-m})$, and ψ-loss $F(m) = 2\min(1, (1 - m)_+)$ with $m_+ = \max(m, 0)$. A number of polynomial-time algorithms with convergence guarantees are readily available for this problem (Ghadermarzy et al., 2018; Wang and Li, 2020; Hong et al., 2020; Alquier et al., 2019). We implement hinge loss (Alquier et al., 2019; Genzel and Stollenwerk, 2020; He et al., 2017) which maintains desirable statistical properties as in 0-1 loss, Here $\mathrm{Risk}_F(\cdot)$ is defined similarly as in (5) with hinge loss in place of 0-1 loss.

Assumption 2 (Assumptions on surrogate loss)**.**

(a) (Approximation error) For any given $\pi \in [-1, 1]$, assume there exist a sequence of tensors $\mathcal{Z}_\pi^{(n)} \in \mathscr{P}_{\mathrm{sgn}}(r)$, such that $\mathrm{Risk}_F(\mathcal{Z}_\pi^{(n)}) - \mathrm{Risk}_F(\bar{\Theta}) \le a_n$, for some sequence $a_n \to 0$ as $n \to \infty$. Furthermore, assume $\|\mathcal{Z}_\pi^{(n)}\|_F \le J$ for some constant $J > 0$.

(b) $F(z) = (1 - z)_+$ is hinge loss.

Assumption 2(a) quantifies the representation capability of and $\mathscr{P}_{\mathrm{sgn}}(r)$. Assumption 2(b) implies the Fisher consistency bound for the weighted risk (Scott, 2011),

$$\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\Theta - \pi) \lesssim \mathrm{Risk}_F(\mathcal{Z}) - \mathrm{Risk}_F(\Theta - \pi), \text{ for all } \pi \in [-1, 1] \text{ and all } \mathcal{Z}.$$

The Fisher consistency enable us to relate the excess risk of the large margin loss to that of 0-1 loss. Under Assumption 2, the resulting estimate from Algorithm 1 enjoys both statistical and computational efficiency.

Theorem 4 (Large-margin estimation)**.** *Consider the same setup as in Theorem 3, and denote* $t_n = \frac{d_{\max} r K \log n}{n}$. *Suppose the surrogate loss F satisfies Assumption 2 with* $a_n \lesssim t_n^{(\alpha+1)/(\alpha+2)}$. *Set* $\lambda \asymp t_n^{(\alpha+1)/(\alpha+2)} + t_n/\rho(\pi, \mathcal{N})$ *in* (46). *Then, with high probability at least* $1 - \exp(-n t_n)$, *we have:*

(a) (Sign tensor estimation). For all $\pi \in [-1, 1]$ except for a finite number of levels,

$$\mathrm{MAE}(\mathrm{sgn}(\hat{\mathcal{Z}}_\pi, \mathrm{sgn}(\Theta - \pi)) \lesssim t_n^{\frac{\alpha}{2+\alpha}} + \frac{1}{\rho^2(\pi, \mathcal{N})} t_n. \tag{10}$$

(b) (Tensor estimation).

$$\mathrm{MAE}(\hat{\Theta}_F, \Theta) \lesssim (t_n \log H)^{\frac{\alpha}{2+\alpha}} + \frac{1+|\mathcal{N}|}{H} + t_n H \log H. \tag{11}$$

In particualr, setting $H \asymp (1+|\mathcal{N}|)^{1/2} t_n^{-1/2}$ *yields the tightest upper bound in* (11).

In principle, uses can choose their own favorite large-margin losses, as long as the base algorithms are sample efficient. The comparison between various large-margin losses has been studied before Bartlett et al. (2006). Note that, instead of using $\hat{\mathcal{Z}}_\pi$ as in existing 1-bit tensor algorithms Ghadermarzy et al. (2018); Wang and Li (2020), we use $\mathrm{sgn}(\hat{\mathcal{Z}}_\pi)$ for more challenging nonparametric estimation. The sign aggregation brings the benefits of flexibility and accuracy over classical low-rank models.

5 Simulations

In this section, we compare our nonparametric tensor method (**NonParaT**) with two alternative approaches: low-rank tensor CP decomposition (**CPT**), and the matrix version of our method applied to tensor unfolding (**NonParaM**). We assess the performance under both complete and incomplete observations. The signal tensors are generated based on four models listed in Table 1. The simulation covers a wide range of complexity, including block tensors, transformed low rank tensors, min/max hypergraphon with logarithm and exponential functions. We consider order-3 tensors of equal dimension $d_1 = d_2 = d_3 = d$, and set $d \in \{15, 20, \ldots, 55, 60\}$, $r = 2$, $H = 10 + (d-15)/5$ in Algorithm 1. For **NonParaM**, we apply Algorithm 1 to each of the three unfolded matrices and report the average error. All summary statistics are averaged across 30 replicates.

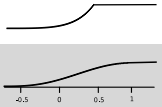
Simulation	Signal Tensor Θ	Rank	Sign Rank	α	$\lvert\mathcal{N}\rvert$	CDF	Noise
1	$\mathcal{C} \times \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \boldsymbol{M}_3$	3^3	$\leq 3^3$	∞	$\leq 3^3$		Uniform $[-0.3, 0.3]$
2	$\lvert\boldsymbol{a} \otimes \mathbf{1} \otimes \mathbf{1} - \mathbf{1} \otimes \boldsymbol{a} \otimes \mathbf{1}\rvert$	d	≤ 3	1	0		Normal $\mathcal{N}(0, 0.15)$
3	$\log(0.5 + Z_{\max})$	$\geq d$	2	∞	d		Uniform $[-0.1, 0.1]$
4	$2.5 - \exp(\mathcal{Z}_{\min}^{1/3})$	$\geq d$	2	∞	d		Normal $\mathcal{N}(0, 0.15)$

Table 1: Simulation models used for comparison. We use $\boldsymbol{M}_k \in \{0,1\}^{d\times 3}$ to denote membership matrices, $\mathcal{C} \in \mathbb{R}^{3\times3\times3}$ the block means, $\boldsymbol{a} = \frac{1}{d}(1, 2, \ldots, d)^T \in \mathbb{R}^d$, $\mathcal{Z}_{\max}$ and $\mathcal{Z}_{\min}$ are order-3 tensors with entries $\frac{1}{d}\max(i,j,k)$ and $\frac{1}{d}\min(i,j,k)$, respectively.

Figure 3 compares the estimation error under full observation. The MAE decreases with tensor dimension for all three methods. We find that our method **NonParaT** achieves the best performance in all scenarios, whereas the second best method is **CPT** for models 1-2, and **NonParaM** for models 3-4. One possible reason is that models 1-2 have controlled multilinear tensor rank, which makes tensor methods **NonParaT** and **CPT** more accurate than matrix methods. For models 3-4, the rank exceeds the tensor dimension, and therefore, the two nonparametric methods **NonParaT** and **NonparaM** exhibit the greater advantage for signal recovery.

Figure 4 shows the completion error against observation fraction. We fix $d = 40$ and gradually increase the observation fraction $\frac{|\Omega|}{d^3}$ from 0.3 to 1. We find that **NonParaT** achieves the lowest error among all methods. Our simulation covers a reasonable range of complexities; for example, model 1 has 3^3 jumps in the CDF of signal Θ, and models 2 and 4 have unbounded noise. Nevertheless, our

method shows good performance in spite of model misspecification. This robustness is appealing in practice because the structure of underlying signal tensor is often unknown.

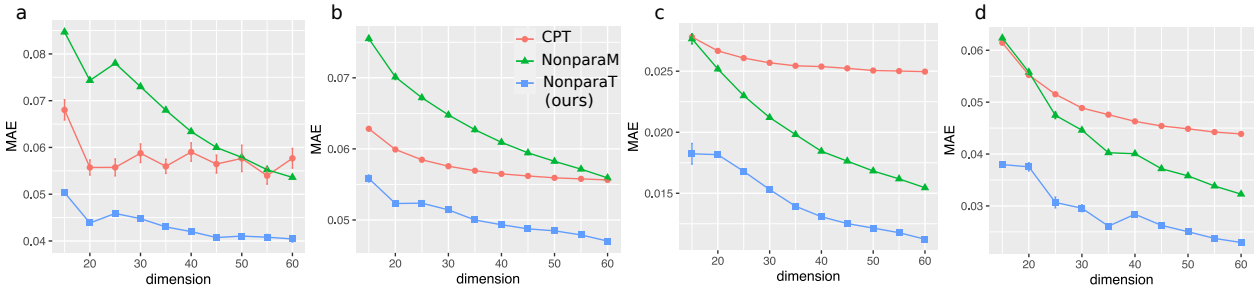


Figure 3: Estimation error versus tensor dimension. Panels (a)-(d) correspond to simulation models 1-4 in Table 1.

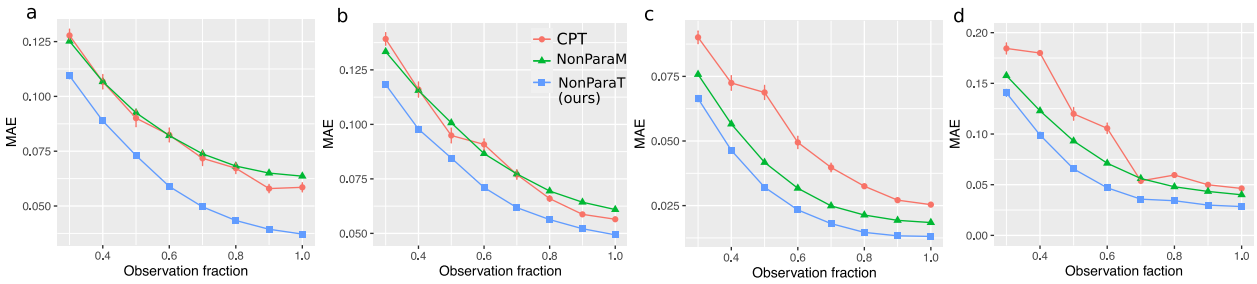


Figure 4: Completion error versus observation fraction. Panels (a)-(d) correspond to simulation models 1-4 in Table 1.

6 Data applications

We apply our method to two tensor datasets, the MRN-114 human brain connectivity data (Wang et al., 2017), and NIPS word occurrence data (Globerson et al., 2007).

6.1 Brain connectivity analysis

The brain dataset records the structural connectivity among 68 brain regions for 114 individuals along with their Intelligence Quotient (IQ) scores. We organize the connectivity data into an order-3 tensor, where entries encode the presence or absence of fiber connections between brain regions across individuals.

Figure 5 shows the MAE based on 5-fold cross-validations with $r = 3, 6, \ldots, 15$ and $H = 20$. We find that our method outperforms CPT in all combinations of ranks and missing rates. The achieved error reduction appears to be more profound as the missing rate increases. This trend highlights the applicability of our method in tensor completion tasks. In addition, our method exhibits a smaller standard error in cross-validation experiments as shown in Figure 5 and Table 2, demonstrating the stability over CPT. One possible reason is that that our estimate is guaranteed to be in $[0, 1]$ (for binary tensor problem where $\mathcal{Y} \in \{0, 1\}^{d_1 \times \cdots d_K}$) whereas CPT estimation may fall outside the valid range $[0, 1]$.

We next investigate the pattern in the estimated signal tensor. Figure 6a shows the identified top edges associated with IQ scores. Specifically, we first obtain a denoised tensor $\hat{\Theta} \in \mathbb{R}^{68 \times 68 \times 114}$ using

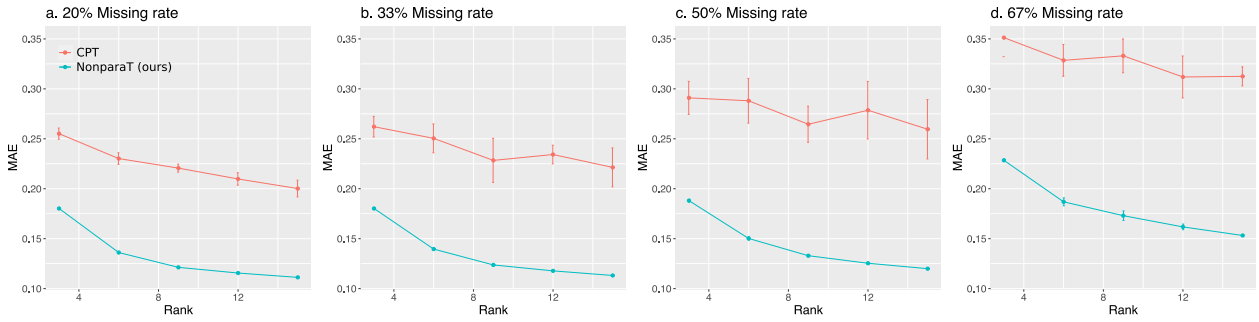


Figure 5: Estimation error versus rank under different missing rate. Panels (a)-(d) correspond to missing rate 20%, 33%, 50%, and 67%, respectively. Error bar represents the standard error over 5-fold cross-validations.

MRN-114 brain connectivity dataset					
Method	$r = 3$	$r = 6$	$r = 9$	$r = 12$	$r = 15$
NonparaT (Ours)	**0.18**(0.001)	**0.14**(0.001)	**0.12**(0.001)	**0.12**(0.001)	**0.11**(0.001)
Low-rank CPT	0.26(0.006)	0.23(0.006)	0.22(0.004)	0.21(0.006)	0.20(0.008)
NIPS word occurrence dataset					
Method	$r = 3$	$r = 6$	$r = 9$	$r = 12$	$r = 15$
NonparaT (Ours)	**0.18**(0.002)	**0.16**(0.002)	**0.15**(0.001)	**0.14**(0.001)	**0.13**(0.001)
Low-rank CPT	0.22(0.004)	0.20(0.007)	0.19(0.007)	0.17(0.007)	0.17(0.007)
Naive imputation (Baseline)	0.32(.001)				

Table 2: MAE comparison in the brain data and NIPS data analysis. Reported MAEs are averaged over five runs of cross-validation, with 20% entries for testing and 80% for training, with standard errors in parentheses. Bold numbers indicate the minimal MAE among three methods. For low-rank CPT, we use R function `rTensor` with default hyperparameters, and for our method, we set $H = 20$.

our method with $r = 10$ and $H = 20$. Then, we perform a regression analysis of $\hat{\Theta}(i, j, :) \in \mathbb{R}^{144}$ against the normalized IQ score across the 144 individuals. The regression model is repeated for each edge $(i, j) \in [68] \times [68]$. We find that top edges represent the interhemispheric connections in the frontal lobes. The result is consistent with recent research on brain connectivity with intelligence (Li et al., 2009; Wang et al., 2017).

6.2 NIPS data analysis

The NIPS dataset consists of word occurrence counts in papers published from 1987 to 2003. We focus on the top 100 authors, 200 most frequent words, and normalize each word count by log transformation with pseudo-count 1. The resulting dataset is an order-3 tensor with entry representing the log counts of words by authors across years.

Table 2 compares the prediction accuracy of different methods. We find that our method substantially outperforms the low-rank CP method for every configuration under consideration. Further increment of rank appears to have little effect on the performance. The comparison highlights the advantage of our method in achieving accuracy while maintaining low complexity. In addition, we also perform naive imputation where the missing values are predicted using the sample average. Both our method and CPT outperform the naive imputation, implying the necessity of incorporating tensor structure in the analysis.

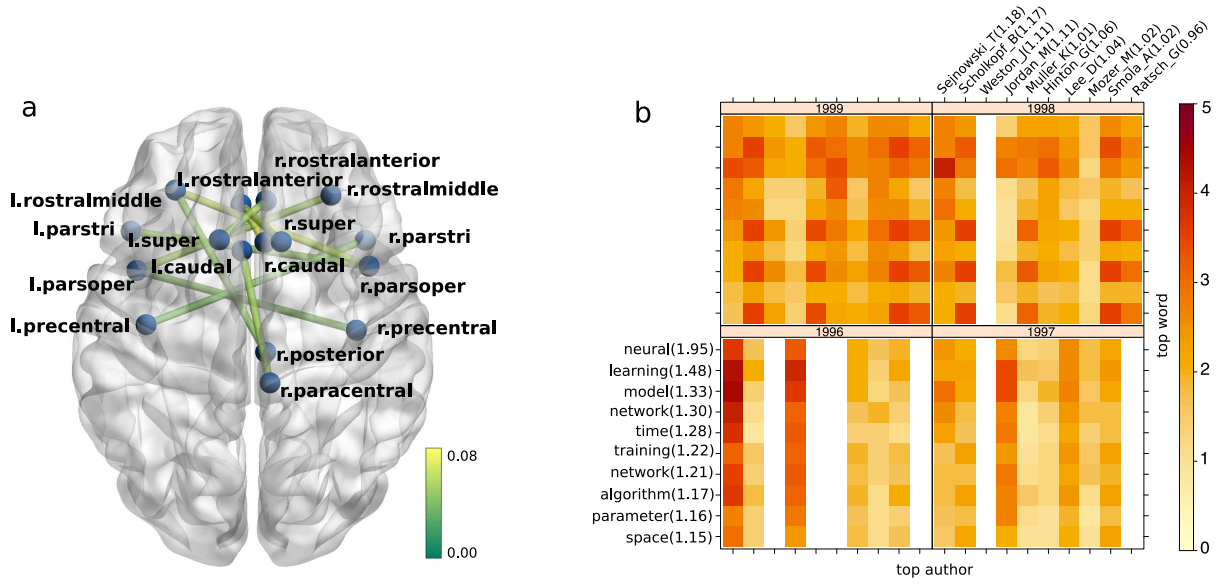


Figure 6: Estimated signal tensors in the data analysis. (a) top edges associated with IQ scores in the brain connectivity data. The color indicates the estimated IQ effect size. (b) top authors and words for years 1996-1999 in the NIPS data. Authors and words are ranked by marginal averages based on $\hat{\Theta}$, where the marginal average is denoted in the parentheses.

We next examine the estimated signal tensor $\hat{\Theta}$ from our method. Figure 6b illustrates the results from NIPS data, where we plot the entries in $\hat{\Theta}$ corresponding to top authors and most-frequent words (after excluding generic words such as *figure*, *results*, etc). The identified pattern is consistent with the active topics in the NIPS publication. Among the top words are *neural* (marginal mean = 1.95), *learning* (1.48), and *network* (1.21), whereas top authors are *T. Sejnowski* (1.18), *B. Scholkopf* (1.17), *M. Jordan* (1.11), and *G. Hinton* (1.06). We also find strong heterogeneity among word occurrences across authors and years. For example, *training* and *algorithm* are popular words for *B. Scholkopf* and *A. Smola* in 1998-1999, whereas *model* occurs more often in *M. Jordan* and in 1996. The detected pattern and achieved accuracy demonstrate the applicability of our method.

7 Additional results and proofs

In this section, we provides additional results not covered in previous sections. Section 7.1 gives detailed explanation to the examples mentioned in Section 1. Section 7.2 supplements Section 3.1 by providing more theoretical results on sign rank and its relationship to tensor rank. Section 7.3 collects the proofs for theorems in the main texts. Lastly, Section 7.4 extends all results to unbounded observations with sub-Gaussian noise.

7.1 Sensitivity of tensor rank to monotonic transformations

In Section 1 of the main paper, we have provided a motivating example to show the sensitivity of tensor rank to monotonic transformations. Here, we describe the details of the example set-up.

The step 1 is to generate a rank-3 tensor $\mathcal{Z}$ based on the CP representation

$$\mathcal{Z} = \boldsymbol{a}^{\otimes 3} + \boldsymbol{b}^{\otimes 3} + \boldsymbol{c}^{\otimes 3},$$

where $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{30}$ are vectors consisting of $N(0,1)$ entries, and the shorthand $\boldsymbol{a}^{\otimes 3} = \boldsymbol{a} \otimes \boldsymbol{a} \otimes \boldsymbol{a}$ denotes the Kronecker power. We then apply $f(z) = (1 + \exp(-cz))^{-1}$ to $\mathcal{Z}$ entrywise, and obtain a transformed tensor $\Theta = f(\mathcal{Z})$.

The step 2 is to determine the rank of Θ. Unlike matrices, the exact rank determination for tensors is NP hard. Therefore, we choose to compute the numerical rank of Θ as an approximation. The numerical rank is determined as the minimal rank for which the relative approximation error is below 0.1, i.e.,

$$\hat{r}(\Theta) = \min \left\{ s \in \mathbb{N}_+ : \min_{\hat{\Theta}:\, \mathrm{rank}(\hat{\Theta}) \le s} \frac{\|\Theta - \hat{\Theta}\|_F}{\|\Theta\|_F} \le 0.1 \right\}.$$

We compute $\hat{r}(\Theta)$ by searching over $s \in \{1, \ldots, 30^2\}$, where for each s, we (approximately) solve the least-square minimization using CP function in R package `rTensor`. We repeat steps 1-2 ten times, and plot the averaged numerical rank of Θ versus transformation level c in Figure 1a.

7.2 Tensor rank and sign-rank

In the main paper, we have provided several tensor examples with high tensor rank but low sign-rank. This section provides more examples and their proofs. Unless otherwise specified, let Θ be an order-K $(d, \ldots, d)$-dimensional tensor.

Example 6 (Structured tensors with repeating entries)**.** Suppose the tensor Θ takes the form

$$\Theta(i_1, \ldots, i_K) = \log\left(1 + \frac{1}{d}\max(i_1, \ldots, i_K)\right), \text{ for all } (i_1, \ldots, i_K) \in [d]^K.$$

Then

$$\mathrm{rank}(\Theta) \ge d, \quad \text{and} \quad \mathrm{srank}(\Theta - \pi) \le 2 \text{ for all } \pi \in \mathbb{R}.$$

Proof of Example 6. We first prove the results for $K = 2$. The full-rankness of Θ is verified from elementary row operations as follows

$$\begin{pmatrix} (\Theta_2 - \Theta_1)/(\log(1+\frac{2}{d}) - \log(1+\frac{1}{d})) \\ (\Theta_3 - \Theta_2)/(\log(1+\frac{3}{d}) - \log(1+\frac{2}{d})) \\ \vdots \\ (\Theta_d - \Theta_{d-1})/(\log(1+\frac{d}{d}) - \log(1+\frac{d-1}{d})) \\ \Theta_d/\log(1+\frac{d}{d}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ddots & \ddots & 0 \\ 1 & 1 & \ddots & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix},$$

where Θ_i denotes the i-th row of Θ. Now it suffices to show $\mathrm{srank}(\Theta - \pi) \le 2$ for π in the feasible range $(\log(1 + \frac{1}{d}),\ \log 2)$. In this case, there exists an index $i^* \in \{2, \ldots, d\}$, such that $\log(1 + \frac{i^*-1}{d}) < \pi \le \log(1 + \frac{i^*}{d})$. By definition, the sign matrix $\mathrm{sgn}(\Theta - \pi)$ takes the form

$$\mathrm{sgn}(\Theta(i,j) - \pi) = \begin{cases} -1, & \text{both } i \text{ and } j \text{ are smaller than } i^*; \\ 1, & \text{otherwise.} \end{cases} \tag{12}$$

Therefore, the matrix $\mathrm{sgn}(\Theta - \pi)$ is a rank-2 block matrix, which implies $\mathrm{srank}(\Theta - \pi) = 2$.

We now extend the results to $K \ge 3$. By definition of the tensor rank, the rank of a tensor is lower bounded by the rank of its matrix slice. So we have $\mathrm{rank}(\Theta) \ge \mathrm{rank}(\Theta(:, :, 1, \ldots, 1)) = d$. For the

sign rank with feasible π, notice that the sign tensor $\operatorname{sgn}(\Theta - \pi)$ takes the similar form as in (12),

$$\operatorname{sgn}(\Theta(i_1, \ldots, i_K) - \pi) = \begin{cases} -1, & i_k < i^* \text{ for all } k \in [K]; \\ 1, & \text{otherwise,} \end{cases} \tag{13}$$

where i^* denotes the index that satisfies $\log(1 + \frac{i^*-1}{d}) < \pi \leq \log(1 + \frac{i^*}{d})$. The equation (13) implies that $\operatorname{sgn}(\Theta - \pi) = -2\boldsymbol{a}^{\otimes K} + 1$, where $\boldsymbol{a} = (1, \ldots, 1, 0, \ldots, 0)^T$ takes 1 on the i-th entry if $i < i^*$ and 0 otherwise. Henceforth $\operatorname{srank}(\Theta - \pi) = 2$. □

In fact, Example 6 is a special case of the following proposition.

Proposition 3 (Structured tensors with repeating entries)**.** *Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuous function such that $g(z) = 0$ has at most $r \geq 1$ distinct real roots. For given numbers $x_{i_k}^{(k)} \in [0, 1]$ for all $i_k \in [d_k]$, define a tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ with entries*

$$\Theta(i_1, \ldots, i_K) = g(\max(x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)})), \quad (i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K]. \tag{14}$$

Then, the sign rank of $(\Theta - \pi)$ satisfies

$$\operatorname{srank}(\Theta - \pi) \leq 2r.$$

The same conclusion holds if we use min *in place of* max *in* (14)*.*

Proof of Proposition 3. We reorder the tensor indices along each mode such that $x_1^{(k)} \leq \cdots \leq x_{d_k}^{(k)}$ for all $k \in [K]$. Based on the construction of $\mathcal{Z}_{\max}$, the reordering does not change the rank of $\mathcal{Z}_{\max}$ or $(\Theta - \pi)$. Let $z_1 < \cdots < z_r$ be the r distinct real roots for the equation $g(z) = \pi$. We separate the proof for two cases, $r = 1$ and $r \geq 2$.

- When $r = 1$. The continuity of $g(\cdot)$ implies that the function $(g(z) - \pi)$ has at most one sign change point. Using similar proof as in Example 6, we have

  $$\operatorname{sgn}(\Theta - \pi) = 1 - 2\boldsymbol{a}^{(1)} \otimes \cdots \otimes \boldsymbol{a}^{(K)} \quad \text{or} \quad \operatorname{sgn}(\Theta - \pi) = 2\boldsymbol{a}^{(1)} \otimes \cdots \otimes \boldsymbol{a}^{(K)} - 1,$$

  where $\boldsymbol{a}^{(k)}$ are binary vectors defined by

  $$\boldsymbol{a}^{(k)} = (\underbrace{1, \ldots, 1,}_{\text{positions for which } x_{i_k}^k < z_1} 0, \ldots, 0)^T, \quad \text{for } k \in [K].$$

  Therefore, $\operatorname{srank}(\Theta - \pi) \leq \operatorname{rank}(\operatorname{sgn}(\Theta - \pi)) = 2$.
- When $r \geq 2$. By continuity, the function $(g(z) - \pi)$ is non-zero and remains an unchanged sign in each of the intervals (z_s, z_{s+1}) for $1 \leq s \leq r - 1$. Define the index set

  $$\mathcal{I} = \{s \in \mathbb{N}_+ \colon \text{the interval } (z_s, z_{s+1}) \text{ in which } g(z) < \pi\}.$$

  We now prove that the sign tensor $\operatorname{sgn}(\Theta - \pi)$ has rank bounded by $2r - 1$. To see this, consider the tensor indices for which $\operatorname{sgn}(\Theta - \pi) = -1$,

  $$\begin{aligned} \{\omega \colon \Theta(\omega) - \pi < 0\} &= \{\omega \colon g(\mathcal{Z}_{\max}(\omega)) < \pi\} \\ &= \cup_{s \in \mathcal{I}} \{\omega \colon \mathcal{Z}_{\max}(\omega) \in (z_s, z_{s+1})\} \end{aligned}$$

$$= \cup_{s\in\mathcal{I}}\Big(\{\omega\colon x_{i_k}^{(k)} < z_{s+1} \text{ for all } k\in[K]\}\cap\{\omega\colon x_{i_k}^{(k)} \le z_s \text{ for all } k\in[K]\}^c\Big). \tag{15}$$

The equation (15) is equivalent to

$$\mathbb{1}(\Theta(i_1,\dots,i_K) < \pi) = \sum_{s\in\mathcal{I}}\left(\prod_k \mathbb{1}(x_{i_k}^{(k)} < z_{s+1}) - \prod_k \mathbb{1}(x_{i_k}^{(k)} \le z_s)\right), \tag{16}$$

for all $(i_1,\dots,i_K)\in[d_1]\times\cdots\times[d_K]$, where $\mathbb{1}(\cdot)\in\{0,1\}$ denotes the indicator function. The equation (16) implies the low-rank representation of $\operatorname{sgn}(\Theta-\pi)$,

$$\operatorname{sgn}(\Theta-\pi) = 1 - 2\sum_{s\in\mathcal{I}}\left(\boldsymbol{a}_{s+1}^{(1)}\otimes\cdots\otimes\boldsymbol{a}_{s+1}^{(K)} - \bar{\boldsymbol{a}}_s^{(1)}\otimes\cdots\otimes\bar{\boldsymbol{a}}_s^{(K)}\right), \tag{17}$$

where $\boldsymbol{a}_{s+1}^{(k)}, \boldsymbol{a}_s^{(k)}$ are binary vectors defined by

$$\boldsymbol{a}_{s+1}^{(k)} = (\underbrace{1,\dots,1,}_{\text{positions for which } x_{i_k}^{(k)} < z_{s+1}} 0,\dots,0)^T, \quad \text{and} \quad \bar{\boldsymbol{a}}_s^{(k)} = (\underbrace{1,\dots,1,}_{\text{positions for which } x_{i_k}^{(k)} \le z_s} 0,\dots,0)^T.$$

Therefore, by (17) and the assumption $|\mathcal{I}|\le r-1$, we conclude that

$$\operatorname{srank}(\Theta-\pi) \le 1 + 2(r-1) = 2r-1.$$

Combining two cases yields that $\operatorname{srank}(\Theta-\pi)\le 2r$ for any $r\ge 1$. □

We next provide several additional examples such that $\operatorname{rank}(\Theta)\ge d$ whereas $\operatorname{srank}(\Theta)\le c$ for a constant c independent of d. We state the examples in the matrix case, i.e, $K=2$. Similar conclusion extends to $K\ge 3$, by the following proposition.

Proposition 4 (Rank relationship between matrices and tensors)**.** *Let $\boldsymbol{M}\in\mathbb{R}^{d_1\times d_2}$ be a matrix. For any given $K\ge 3$, define an order-K tensor $\Theta\in\mathbb{R}^{d_1\times\cdots\times d_K}$ by*

$$\Theta = \boldsymbol{M}\otimes\mathbf{1}_{d_3}\otimes\cdots\otimes\mathbf{1}_{d_K},$$

where $\mathbf{1}_{d_k}\in\mathbb{R}^{d_k}$ denotes an all-one vector, for $3\le k\le K$. Then we have

$$\operatorname{rank}(\Theta) = \operatorname{rank}(\boldsymbol{M}), \quad \textit{and} \quad \operatorname{srank}(\Theta-\pi) = \operatorname{srank}(\boldsymbol{M}-\pi) \textit{ for all } \pi\in\mathbb{R}.$$

Proof of Proposition 4. The conclusion directly follows from the definition of tensor rank. □

Example 7 (Stacked banded matrices)**.** Let $\boldsymbol{a} = (1,2,\dots,d)^T$ be a d-dimensional vector, and define a d-by-d banded matrix $\boldsymbol{M} = |\boldsymbol{a}\otimes\mathbf{1} - \mathbf{1}\otimes\boldsymbol{a}|$. Then

$$\operatorname{rank}(\boldsymbol{M}) = d, \quad \text{and} \quad \operatorname{srank}(\boldsymbol{M}-\pi)\le 3, \quad \text{for all } \pi\in\mathbb{R}.$$

Proof of Example 7. Note that $\boldsymbol{M}$ is a banded matrix with entries

$$\boldsymbol{M}(i,j) = |i-j|, \quad \text{for all } (i,j)\in[d]^2.$$

Elementary row operation shows that $\boldsymbol{M}$ is full rank as follows,

$$\begin{pmatrix} (\boldsymbol{M}_1+\boldsymbol{M}_d)/(d-1) \\ \boldsymbol{M}_1-\boldsymbol{M}_2 \\ \boldsymbol{M}_2-\boldsymbol{M}_3 \\ \vdots \\ \boldsymbol{M}_{d-1}-\boldsymbol{M}_d \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ -1 & 1 & 1 & \cdots & 1 & 1 \\ -1 & -1 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & -1 & 1 \end{pmatrix}.$$

We now show $\operatorname{srank}(\boldsymbol{M}-\pi) \le 3$ by construction. Define two vectors $\boldsymbol{b} = (2^{-1}, 2^{-2}, \ldots, 2^{-d})^T \in \mathbb{R}^d$ and $\operatorname{rev}(\boldsymbol{b}) = (2^{-d}, \ldots, 2^{-1})^T \in \mathbb{R}^d$. We construct the following matrix

$$\boldsymbol{A} = \boldsymbol{b} \otimes \operatorname{rev}(\boldsymbol{b}) + \operatorname{rev}(\boldsymbol{b}) \otimes \boldsymbol{b}. \tag{18}$$

The matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is banded with entries

$$\boldsymbol{A}(i,j) = \boldsymbol{A}(j,i) = \boldsymbol{A}(d-i,d-j) = \boldsymbol{A}(d-j,d-i) = 2^{-d-1}\left(2^{j-i}+2^{i-j}\right), \text{ for all } (i,j) \in [d]^2.$$

Furthermore, the entry value $\boldsymbol{A}(i,j)$ decreases with respect to $|i-j|$; i.e.,

$$\boldsymbol{A}(i,j) \ge \boldsymbol{A}(i',j'), \quad \text{for all } |i-j| \ge |i'-j'|. \tag{19}$$

Notice that for a given $\pi \in \mathbb{R}$, there exists $\pi' \in \mathbb{R}$ such that $\operatorname{sgn}(\boldsymbol{A}-\pi') = \operatorname{sgn}(\boldsymbol{M}-\pi)$. This is because both $\boldsymbol{A}$ and $\boldsymbol{M}$ are banded matrices satisfying monotonicity (19). By definition (18), $\boldsymbol{A}$ is a rank-2 matrix. Henceforce, $\operatorname{srank}(\boldsymbol{M}-\pi) = \operatorname{srank}(\boldsymbol{A}-\pi') \le 3$. □

Remark 1. The tensor analogy of banded matrices $\Theta = |\boldsymbol{a}\otimes\mathbf{1}\otimes\mathbf{1} - \mathbf{1}\otimes\boldsymbol{a}\otimes\mathbf{1}|$ is used as simulation model 3 in the main paper.

Example 8 (Stacked identity matrices)**.** Let $\boldsymbol{I}$ be a d-by-d identity matrix. Then

$$\operatorname{rank}(\boldsymbol{I}) = d, \quad \text{and} \quad \operatorname{srank}(\boldsymbol{I}-\pi) \le 3 \text{ for all } \pi \in \mathbb{R}.$$

Proof of Proposition 8. Depending on the value of π, the sign matrix $\operatorname{sgn}(\boldsymbol{I}-\pi)$ falls into one of the two cases:

(a) $\operatorname{sgn}(\boldsymbol{I}-\pi)$ is a matrix of all 1, or of all -1;

(b) $\operatorname{sgn}(\boldsymbol{I}-\pi) = 2\boldsymbol{I} - \mathbf{1}_d \otimes \mathbf{1}_d$.

The first cases are trivial, so it suffices to show $\operatorname{srank}(\boldsymbol{I}-\pi) \le 3$ in the third case.

Based on Example 7, the rank-2 matrix $\boldsymbol{A}$ in (18) satisfies

$$\boldsymbol{A}(i,j)\begin{cases} = 2^{-d}, & i = j, \\ \ge 2^{-d} + 2^{-d-2}, & i \ne j. \end{cases}$$

Therefore, $\operatorname{sgn}\left(2^{-d}+2^{-d-3}-\boldsymbol{A}\right) = 2\boldsymbol{I} - \mathbf{1}_d \otimes \mathbf{1}_d$. We conclude that $\operatorname{srank}(\boldsymbol{I}-\pi) \le \operatorname{rank}(2^{-d} + 2^{-d-3} - \boldsymbol{A}) = 3$. □

7.3 Proofs

7.3.1 Proofs of Propositions 1-2

Proof of Proposition 1.

Part (a). The strictly monotonicity of g implies that the inverse function $g^{-1}\colon \mathbb{R} \to \mathbb{R}$ is well-defined. When g is strictly increasing, the mapping $x \mapsto g(x)$ is sign preserving. Specifically, if $x \geq 0$, then $g(x) \geq g(0) = 0$. Conversely, if $g(x) \geq 0 = g(0)$, then applying g^{-1} to both sides gives $x \geq 0$. When g is strictly decreasing, the mapping $x \mapsto g(x)$ is sign reversing. Specifically, if $x \geq 0$, then $g(x) \leq g(0) = 0$. Conversely, if $g(x) \geq 0 = g(0)$, then applying g^{-1} to both sides gives $x \leq 0$. Therefore, $\Theta \simeq g(\Theta)$, or $\Theta \simeq -g(\Theta)$. Since constant multiplication does not change the tensor rank, we have $\operatorname{srank}(\Theta) = \operatorname{srank}(g(\Theta)) \leq \operatorname{rank}(g(\Theta))$.

Part (b). See Section 7.2 for constructive examples.

□

Proof of Proposition 2. Fix $\pi \in [-1, 1]$. Based on the definition of classification loss $L(\cdot, \cdot)$, the function $\operatorname{Risk}(\cdot)$ relies only on the sign pattern of the tensor. Therefore, without loss of generality, we assume both $\bar{\Theta}, \mathcal{Z} \in \{-1, 1\}^{d_1 \times \cdots \times d_K}$ are binary tensors. We evaluate the excess risk

$$\operatorname{Risk}(\mathcal{Z}) - \operatorname{Risk}(\bar{\Theta}) = \mathbb{E}_{\omega \sim \Pi} \underbrace{\mathbb{E}_{\mathcal{Y}(\omega)} \left\{ |\mathcal{Y}(\omega) - \pi| \left[\left| \mathcal{Z}(\omega) - \operatorname{sgn}(\bar{\mathcal{Y}}(\omega)) \right| - \left| \bar{\Theta}(\omega) - \operatorname{sgn}(\bar{\mathcal{Y}}(\omega)) \right| \right] \right\}}_{\stackrel{\text{def}}{=} I(\omega)}. \quad (20)$$

Denote $y = \mathcal{Y}(\omega)$, $z = \mathcal{Z}(\omega)$, $\bar{\theta} = \bar{\Theta}(\omega)$, and $\theta = \Theta(\omega)$. The expression of $I(\omega)$ is simplified as

$$\begin{aligned} I(\omega) &= \mathbb{E}_{y|\omega} \left[(y - \pi)(\bar{\theta} - z) \mathbb{1}(y \geq \pi) + (\pi - y)(z - \bar{\theta}) \mathbb{1}(y < \pi) \right] \\ &= \mathbb{E}_{y|\omega} \left[(\bar{\theta} - z)(y - \pi) \right] \\ &= \left[\operatorname{sgn}(\theta - \pi) - z \right] (\theta - \pi) \\ &= |\operatorname{sgn}(\theta - \pi) - z| |\theta - \pi| \geq 0, \end{aligned} \quad (21)$$

where the third line uses the fact $\mathbb{E}y = \theta$ and $\bar{\theta} = \operatorname{sgn}(\theta - \pi)$, and the last line uses the assumption $z \in \{-1, 1\}$. The equality (21) is attained when $z = \operatorname{sgn}(\theta - \pi)$ or $\theta = \pi$. Combining (21) with (20), we conclude that, for all $\mathcal{Z} \in \{-1, 1\}^{d_1 \times \cdots \times d_K}$,

$$\operatorname{Risk}(\mathcal{Z}) - \operatorname{Risk}(\bar{\Theta}) = \mathbb{E}_{\omega \sim \Pi} |\operatorname{sgn}(\Theta(\omega) - \pi) - \mathcal{Z}(\omega)| |\Theta(\omega) - \pi| \geq 0. \quad (22)$$

In particular, setting $\mathcal{Z} = \bar{\Theta} = \operatorname{sgn}(\Theta - \pi)$ in (22) yields the minimum. Therefore,

$$\operatorname{Risk}(\bar{\Theta}) = \min\{\operatorname{Risk}(\mathcal{Z})\colon \mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}\} \leq \min\{\operatorname{Risk}(\mathcal{Z})\colon \operatorname{rank}(\mathcal{Z}) \leq r\}.$$

Since $\operatorname{srank}(\Theta - \pi) \leq r$ by assumption, the last inequality becomes equality. The proof is complete.

□

7.3.2 Proof of Theorem 1

Proof of Theorem 1. Fix $\pi \notin \mathcal{N}$. Based on (22) in Proposition 2, we have

$$\operatorname{Risk}(\mathcal{Z}) - \operatorname{Risk}(\bar{\Theta}) = \mathbb{E} \left[|\operatorname{sgn}\mathcal{Z} - \operatorname{sgn}\bar{\Theta}| |\bar{\Theta}| \right]. \quad (23)$$

The Assumption 1 states that

$$\mathbb{P}\left(|\bar{\Theta}| \leq t\right) \leq \begin{cases} ct^{\alpha}, & \text{for all } \Delta s \leq t < \rho(\pi, \mathcal{N}), \\ C\Delta s, & \text{for all } 0 \leq t < \Delta s. \end{cases} \tag{24}$$

Without further specification, all relevant probability statements, such as $\mathbb{E}$ and $\mathbb{P}$, are with respect to $\omega \sim \Pi$.

We divide the proof into two cases: $\alpha > 0$ and $\alpha = \infty$.

- Case 1: $\alpha > 0$.

  By (23), for all $0 \leq t < \rho(\pi, \mathcal{N})$,

$$\begin{aligned} \text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) &\geq t\mathbb{E}\left(|\text{sgn}\mathcal{Z} - \text{sgn}\bar{\Theta}|\mathbb{1}\{|\bar{\Theta}| > t\}\right) \\ &\geq 2t\mathbb{P}\left(\text{sgn}\mathcal{Z} \neq \text{sgn}\bar{\Theta} \text{ and } |\bar{\Theta}| > t\right) \\ &\geq 2t\Big\{\mathbb{P}\left(\text{sgn}\mathcal{Z} \neq \text{sgn}\bar{\Theta}\right) - \mathbb{P}\left(|\bar{\Theta}| \leq t\right)\Big\} \\ &\geq t\Big\{\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) - C\Delta s - 2ct^{\alpha}\Big\}, \end{aligned} \tag{25}$$

  where the last line follows from the definition of MAE and (24). We maximize the lower bound (25) with respect to t, and obtain the optimal t_{opt},

$$t_{\text{opt}} = \begin{cases} \rho(\pi, \mathcal{N}), & \text{if MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) > \text{cut-off}, \\ \left[\frac{1}{2c(1+\alpha)}(\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) - C\Delta S)\right]^{1/\alpha}, & \text{if MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) \leq \text{cut-off}. \end{cases}$$

  where we have denoted the cut-off $= 2c(1+\alpha)\rho^{\alpha}(\pi, \mathcal{N}) + C\Delta s$. The corresponding lower bound of the inequality (25) becomes

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) \geq \begin{cases} c_1\rho(\pi, \mathcal{N})\left[\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) - C\Delta s\right], & \text{if MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) > \text{cut-off}, \\ c_2\left[\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) - C\Delta s\right]^{\frac{1+\alpha}{\alpha}}, & \text{if MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) \leq \text{cut-off}, \end{cases}$$

  where $c_1, c_2 > 0$ are two constants independent of $\mathcal{Z}$. Combining both cases gives

$$\begin{aligned} \text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) &\lesssim [\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})}\left[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})\right] + \Delta s \\ &\leq C(\pi)[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})]^{\frac{\alpha}{1+\alpha}} + \Delta s, \end{aligned}$$

  where $C(\pi) > 0$ is a multiplicative factor independent of $\mathcal{Z}$.
- Case 2: $\alpha = \infty$. The inequality (25) now becomes

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) \geq t\left[\text{MAE}(\text{sgn}\bar{\Theta}, \text{sgn}\mathcal{Z}) - C\Delta s\right], \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}). \tag{26}$$

  The conclusion follows by taking $t = \frac{\rho(\pi, \mathcal{N})}{2}$ in the inequality (26).

□

Remark 2. The proof of Theorem 1 shows that, under global α-smoothness of Θ,

$$\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) \lesssim [\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})}\left[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})\right] + \Delta s, \tag{27}$$

for all $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$. For fixed π, the second term is absorbed into the first term.

7.3.3 Proof of Theorem 2

The following lemma provides the variance-to-mean relationship implied by the α-smoothness of Θ. The relationship plays a key role in determining the convergence rate based on empirical process theory (Shen and Wong, 1994); also see Theorem 5.

Lemma 1 (Variance-to-mean relationship)**.** *Consider the same setup as in Theorem 2. Fix $\pi \notin \mathcal{N}$. Let $L(\mathcal{Z}, \bar{Y}_\Omega)$ be the π-weighted classification loss*

$$\begin{aligned} L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) &= \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \underbrace{|\bar{\mathcal{Y}}(\omega)|}_{weight} \times \underbrace{|\mathrm{sgn}\mathcal{Z}(\omega) - \mathrm{sgn}\bar{\mathcal{Y}}(\omega)|}_{classification\ loss} \\ &= \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}), \end{aligned} \tag{28}$$

where we have denoted the function $\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}) \stackrel{def}{=} |\bar{\mathcal{Y}}(\omega)||\mathrm{sgn}\mathcal{Z}(\omega) - \mathrm{sgn}\bar{\mathcal{Y}}(\omega)|$. Under Assumption 1 of the α-smoothness of Θ, we have

$$\mathrm{Var}[\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)] \lesssim [\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})}[\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})] + \Delta s, \tag{29}$$

for all tensors $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$. Here the expectation and variance are taken with respect to both $\mathcal{Y}$ and $\omega \sim \Pi$.

Proof of Lemma 1. We expand the variance by

$$\begin{aligned} \mathrm{Var}[\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)] &\lesssim \mathbb{E}|\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)|^2 \\ &\lesssim \mathbb{E}|\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)| \\ &\le \mathbb{E}|\mathrm{sgn}\mathcal{Z} - \mathrm{sgn}\bar{\Theta}| = \mathrm{MAE}(\mathrm{sgn}\mathcal{Z}, \mathrm{sgn}\bar{\Theta}), \end{aligned} \tag{30}$$

where the second line comes from the boundedness of classification loss $L(\cdot, \cdot)$, and the third line comes from the inequality $||a - b| - |c - b|| \le |a - b|$ for $a, b, c \in \{-1, 1\}$, together with the boundedness of classification weight $|\bar{\mathcal{Y}}(\omega)|$. Here we have absorbed the constant multipliers in $\lesssim$. The conclusion (29) then directly follows by applying Remark 2 to (30). □

Proof of Theorem 2. Fix $\pi \notin \mathcal{N}$. For notational simplicity, we suppress the subscript π and write $\hat{\mathcal{Z}}$ in place of $\hat{\mathcal{Z}}_\pi$. Denote $n = |\Omega|$ and $\rho = \rho(\pi, \mathcal{N})$.

Because the classification loss $L(\cdot, \cdot)$ is scale-free, i.e., $L(\mathcal{Z}, \cdot) = L(c\mathcal{Z}, \cdot)$ for every $c > 0$, we consider the estimation subject to $\|\mathcal{Z}\|_F \le 1$ without loss of generality. Specifically, let

$$\hat{\mathcal{Z}} = \underset{\mathcal{Z}:\ \mathrm{rank}(\mathcal{Z}) \le r, \|\mathcal{Z}\|_F \le 1}{\arg\min} L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega). \tag{31}$$

We next apply the empirical process theory to bound $\hat{\mathcal{Z}}$. To facilitate the analysis, we view the data $\bar{\mathcal{Y}}_\Omega = \{\bar{\mathcal{Y}}(\omega) \colon \omega \in \Omega\}$ as a collection of n independent random variables where the randomness is from both $\bar{\mathcal{Y}}$ and $\omega \sim \Pi$. Write the index set $\Omega = \{1, \ldots, n\}$, so the loss function (28) becomes

$$L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(\mathcal{Z}, \bar{\mathcal{Y}}).$$

We use $f_{\mathcal{Z}}\colon [d_1]\times\cdots\times[d_n]\to\mathbb{R}$ to denote the function induced by tensor $\mathcal{Z}$ such that $f_{\mathcal{Z}}(\omega)=\mathcal{Z}(\omega)$ for $\omega\in[d_1]\times\cdots\times[d_K]$. Under this set-up, the quantity of interest

$$L(\mathcal{Z},\bar{\mathcal{Y}}_\Omega)-L(\bar{\Theta},\bar{\mathcal{Y}}_\Omega)=\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left[\ell_i(\mathcal{Z},\bar{\mathcal{Y}})-\ell_i(\bar{\Theta},\bar{\mathcal{Y}})\right]}_{\stackrel{\text{def}}{=}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})},$$

is an empirical process induced by function $f_{\mathcal{Z}}\in\mathcal{F}_{\mathcal{T}}$ where $\mathcal{T}=\{\mathcal{Z}\colon \operatorname{rank}(\mathcal{Z})\le r,\ \|\mathcal{Z}\|_F\le 1\}$. Note that there is an one-to-one correspondence between sets $\mathcal{F}_{\mathcal{T}}$ and $\mathcal{T}$.

Let L_n denote the desired convergence rate to seek. By definition of $\hat{\mathcal{Z}}$ in (31), we have,

$$L(\hat{\mathcal{Z}},\bar{\mathcal{Y}}_\Omega)-L(\bar{\Theta},\bar{\mathcal{Y}}_\Omega)=\frac{1}{n}\sum_{i=1}^{n}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})\le 0.$$

Therefore, we have the following inclusion of probability events,

$$\begin{aligned}
&\left\{(\omega,\mathcal{Y}_\omega)\colon \operatorname{Risk}(\hat{\mathcal{Z}})-\operatorname{Risk}(\bar{\Theta})\ge L_n\right\}\\
&\subset\left\{(\omega,\mathcal{Y}_\omega)\colon \exists\mathcal{Z}\text{ s.t. }\operatorname{rank}(\mathcal{Z})\le r,\operatorname{Risk}(\mathcal{Z})-\operatorname{Risk}(\bar{\Theta})\ge L_n,\text{ and }\frac{1}{n}\sum_{i=1}^{n}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})\le 0\right\}\\
&\subset\left\{(\omega,\mathcal{Y}_\omega)\colon \sup_{\substack{\operatorname{rank}(\mathcal{Z})\le r\\ \operatorname{Risk}(\mathcal{Z})-\operatorname{Risk}(\bar{\Theta})\ge L_n}} -\frac{1}{n}\sum_{i=1}^{n}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})\ge 0\right\}\\
&\subset\bigcup_{\ell=1}^{\infty}\left\{(\omega,\mathcal{Y}_\omega)\colon \sup_{\mathcal{Z}\in A_\ell} -\frac{1}{n}\sum_{i=1}^{n}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})\ge 0\right\},
\end{aligned}\tag{32}$$

where we have partitioned $\{\mathcal{Z}\colon \operatorname{rank}(\mathcal{Z})\le r \text{ and } \operatorname{Risk}(\mathcal{Z})-\operatorname{Risk}(\bar{\Theta})\ge L_n\}$ in to union of A_ℓ with

$$A_\ell=\{\mathcal{Z}\colon \operatorname{rank}(\mathcal{Z})\le r \text{ and } \ell L_n\le\operatorname{Risk}(\mathcal{Z})-\operatorname{Risk}(\bar{\Theta})<(\ell+1)L_n\},$$

for $\ell=1,2,\ldots$. Let Γ denote the target probability for the first line in (32). To bound Γ, we bound the sum of probability over the sets A_ℓ. For each A_ℓ, we consider the centered empirical process,

$$v_n(f_{\mathcal{Z}}):=-\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_i(f_{\mathcal{Z}},\bar{\Theta})-\mathbb{E}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})\right).\tag{33}$$

Notice $(\ell+1)L_n\ge\mathbb{E}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})=\operatorname{Risk}(\mathcal{Z})-\operatorname{Risk}(\bar{\Theta})\ge\ell L_n$ for all $\mathcal{Z}\in A_\ell$. Combining (32), (33) and union bound yields

$$\Gamma\le\sum_{\ell=1}^{\infty}\mathbb{P}\left\{\sup_{\mathcal{Z}\in A_\ell}v_n(f_{\mathcal{Z}})\ge\ell L_n=:M(\ell)\right\}.\tag{34}$$

Notice that, based on Lemma 1, the variance of empirical process is bounded by

$$\begin{aligned}
\sup_{\mathcal{Z}\in A_\ell}\operatorname{Var}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})&\lesssim\sup_{\mathcal{Z}\in A_\ell}\left(\left[\mathbb{E}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})\right]^{\frac{\alpha}{1+\alpha}}+\frac{1}{\rho}\mathbb{E}\Delta_i(f_{\mathcal{Z}},\bar{\Theta})\right)+\Delta s\\
&\le M(\ell+1)^{\frac{\alpha}{1+\alpha}}+\frac{1}{\rho}M(\ell+1)+\Delta s=:V(\ell).
\end{aligned}$$

We next bound the right-hand side of (34) by choosing L_n that satisfies conditions in Theorem 5 (The specification of L_n is deferred to the next paragraph). One such L_n is chosen, Theorem 5 gives us

$$\begin{aligned}\Gamma &\lesssim \sum_{\ell=1}^{\infty} \exp\left(-\frac{nM^2(\ell)}{V(\ell)+2M(\ell)}\right) \\ &\lesssim \sum_{\ell=1}^{\infty} \exp(-\rho\ell n L_n) \\ &\leq \left(\frac{e^{-n\rho L_n}}{1-e^{-n\rho L_n}}\right).\end{aligned} \tag{35}$$

Now, we specify L_n that satisfies the condition of Theorem 5. The quantity L_n is determined by the solution to the following inequality,

$$\sup_{\ell\geq 1}\frac{1}{x}\int_{x}^{\sqrt{x^{\alpha/(\alpha+1)}+x/\rho+\Delta s}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon,\mathcal{F}_{\mathcal{T}},\|\cdot\|_2)}d\varepsilon \lesssim n^{1/2}, \quad \text{where } x=\ell L_n. \tag{36}$$

In particular, the smallest L_n satisfying (36) yields the best upper bound of the error rate. Here $\mathcal{H}_{[\,]}(\varepsilon,\mathcal{F}_{\mathcal{T}},\|\cdot\|_2)$ denotes the L_2-norm, ε-bracketing number (c.f. Definition 3) for function family $\mathcal{F}_{\mathcal{T}}$.

Based on Lemma 2, the inequality (36) is satisfied with the choice

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)}+\frac{t_n}{\rho}, \quad \text{where } t_n=\left(\frac{d_{\max} rK\log n}{n}\right) \text{ and } d_{\max}:=\max_{k\in[K]} d_k.$$

Finally, it follows from Theorem 5 and (35) that

$$\begin{aligned}\mathbb{P}\left\{\text{Risk}(\hat{\mathcal{Z}})-\text{Risk}(\bar{\Theta})\geq L_n\right\} &\lesssim \left(\frac{e^{-n\rho L_n}}{1-e^{-n\rho L_n}}\right) \\ &\lesssim e^{-nt_n},\end{aligned}$$

where the last inequality uses the fact that $\rho L_n \gtrsim t_n \gtrsim \frac{1}{n}$ by our choice of L_n and t_n.

Inserting the above bound into (27) gives that, with high probability at least $1-\exp(-nt_n)$,

$$\begin{aligned}\text{MAE}(\text{sgn}\hat{\mathcal{Z}},\text{sgn}\bar{\Theta}) &\lesssim [\text{Risk}(\hat{\mathcal{Z}})-\text{Risk}(\bar{\Theta})]^{\alpha/(\alpha+1)}+\frac{1}{\rho}[\text{Risk}(\hat{\mathcal{Z}})-\text{Risk}(\bar{\Theta})]+\Delta s \\ &\lesssim t_n^{\alpha/(\alpha+2)}+\frac{1}{\rho^{\alpha/\alpha+1}}t_n^{\alpha/(\alpha+1)}+\frac{1}{\rho}t_n^{(\alpha+1)/(\alpha+2)}+\frac{1}{\rho^2}t_n \\ &\leq 4t_n^{\alpha/(\alpha+2)}+\frac{4}{\rho^2}t_n,\end{aligned} \tag{37}$$

where the second line uses the fact that $\Delta s \ll t_n$, and the last line follows from the fact that $a(b^2+b^{(\alpha+2)/(\alpha+1)}+b+1)\leq 4a(b^2+1)$ with $a=\frac{t_n}{\rho^2}$ and $b=\rho t_n^{-1/(\alpha+2)}$. We plug t_n into (37) and absorb the term K into the constant. The conclusion is then proved by noting $n=|\Omega|$ by definition. □

Definition 3 (Bracketing number)**.** Consider a family of functions $\mathcal{F}$, and let $\varepsilon > 0$. Let $\mathcal{X}$ denote the domain space equipped with measure Π. We call $\{(f_m^l, f_m^u)\}_{m=1}^M$ an L_2-metric, ε-bracketing function set of $\mathcal{F}$, if for every $f \in \mathcal{F}$, there exists an $m \in [M]$ such that

$$f_m^l(x) \le f(x) \le f_m^u(x), \quad \text{for all } x \in \mathcal{X},$$

and

$$\|f_m^l - f_m^u\|_2 \stackrel{\text{def}}{=} \sqrt{\mathbb{E}_{x\sim\Pi}|f_m^l(x) - f_m^u(x)|^2} \le \varepsilon, \text{ for all } m = 1, \dots, M.$$

The bracketing number with L_2-metric, denoted $\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_2)$, is the logarithm of the smallest cardinality of the ε-bracketing function set of $\mathcal{F}$.

Lemma 2 (Bracketing complexity of low-rank tensors)**.** *Define the family of rank-r bounded tensors $\mathcal{T} = \{\mathcal{Z} \in \mathbb{R}^{d_1\times\cdots\times d_K} : \operatorname{rank}(\mathcal{Z}) \le r,\ \|\mathcal{Z}\|_F \le 1\}$ and the induced function family $\mathcal{F}_\mathcal{T} = \{f_\mathcal{Z} : \mathcal{Z} \in \mathcal{T}\}$. Set*

$$L_n \asymp \left(\frac{d_{\max} r K \log n}{n}\right)^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi, \mathcal{N})}\left(\frac{d_{\max} r K \log n}{n}\right), \text{ where } d_{\max} = \max_{k\in[K]} d_k.$$

Then, the following inequality is satisfied provided that $\Delta s \lesssim n^{-1}$,

$$\sup_{\ell\ge 1} \frac{1}{\ell L_n} \int_{\ell L_n}^{\sqrt{\ell L_n^{\alpha/(\alpha+1)} + \frac{\ell L_n}{\rho(\pi,\mathcal{N})} + \Delta s}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}_\mathcal{T}, \|\cdot\|_2)}\, d\varepsilon \le C n^{1/2}, \tag{38}$$

where $C > 0$ is a constant independent of r, K and d_{max}.

Proof of Lemma 2. To simplify the notation, we denote $\rho = \rho(\pi, \mathcal{N})$. Notice that

$$\|f_{\mathcal{Z}_1} - f_{\mathcal{Z}_1}\|_2 \le \|f_{\mathcal{Z}_1} - f_{\mathcal{Z}_1}\|_\infty \le \|\mathcal{Z}_1 - \mathcal{Z}_1\|_F \quad \text{for all } \mathcal{Z}_1, \mathcal{Z}_2 \in \mathcal{T}.$$

It follows from Kosorok (2007, Theorem 9.22) that the L_2-metric, (2ϵ)-bracketing number of $\mathcal{F}_\mathcal{T}$ is bounded by

$$\mathcal{H}_{[\,]}(2\varepsilon, \mathcal{F}_\mathcal{T}, \|\cdot\|_2) \le \mathcal{H}(\varepsilon, \mathcal{T}, \|\cdot\|_F) \le C d_{\max} r K \log \frac{K}{\varepsilon}.$$

The last inequality is from the covering number bounds for rank-r bounded tensors; see Mu et al. (2014, Lemma 3). Inserting the bracketing number into (38) gives

$$g(L, \ell) = \frac{1}{\ell L} \int_{\ell L}^{\sqrt{\ell L^{\alpha/(\alpha+1)} + \rho^{-1}\ell L + \Delta s}} \sqrt{d_{\max} r K \log\left(\frac{K}{\varepsilon}\right)}\, d\varepsilon. \tag{39}$$

Define $g(L) := \sup_{\ell\ge 1} g(L, \ell)$. By the monotonicity the integrand in (39), we bound $g(L)$ by

$$\begin{aligned}
g(L) &\le \sup_{\ell\ge 1} \frac{\sqrt{d_{\max} r K}}{\ell L} \int_{\ell L}^{\sqrt{\ell L^{\alpha/(\alpha+1)} + \rho^{-1}\ell L + n^{-1}}} \sqrt{\log\left(\frac{K}{\ell L}\right)}\, d\varepsilon \\
&\le \sup_{\ell\ge 1} \sqrt{d_{\max} r K \log\left(\frac{K}{\ell L}\right)} \left(\frac{(\ell L)^{\alpha/(2\alpha+2)} + \sqrt{\rho^{-1}\ell L + n^{-1}}}{\ell L} - 1\right) \\
&\lesssim \sqrt{d_{\max} r K \log(1/L)} \left[\frac{1}{L^{(\alpha+2)/(2\alpha+2)}} + \frac{1}{\sqrt{\rho L}}\left(1 + \frac{\rho}{2nL}\right)\right],
\end{aligned} \tag{40}$$

where the the second line follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b > 0$ and the last line comes from the fact that the bound achieves maximum when $\ell = 1$. It remains to verify that $g(L_n) \le Cn^{1/2}$ for L_n specified in (38). Plugging L_n into the last line of (40) gives

$$\begin{aligned} g(L_n) &\le \sqrt{d_{\max} r K \log(1/L_n)} \left(\frac{1}{L_n^{(\alpha+2)/(2\alpha+2)}} + \frac{2}{\sqrt{\rho L_n}} \right) \\ &\le \sqrt{d_{\max} r K \log n} \left(\left[\left(\frac{d_{\max} r K \log n}{n} \right)^{\frac{\alpha+1}{\alpha+2}} \right]^{-\frac{\alpha+2}{2\alpha+2}} + \left[2\rho \left(\frac{d_{\max} r K \log n}{\rho n} \right) \right]^{-\frac{1}{2}} \right) \\ &\le Cn^{1/2}, \end{aligned}$$

where $C > 0$ is a constant independent of r, K and $d_{\max}$. The proof is therefore complete. □

Theorem 5 (Theorem 3 in Shen and Wong (1994))**.** *Let $\mathcal{F}$ be a class of functions defined on $\mathcal{X}$ with $\sup_{f\in\mathcal{F}} \|f\|_\infty \le T$. Let $(\boldsymbol{X}_i)_{i=1}^n$ be i.i.d. random variables with distribution $\mathbb{P}_{\boldsymbol{X}}$ over $\mathcal{X}$. Set $\sup_{f\in\mathcal{F}} \mathrm{Var} f(\boldsymbol{X}) = V < \infty$. Define the empirical process $\hat{\mathbb{E}} f = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i)$. Define x_n^* to be the solution to the following inequality*

$$\frac{1}{x} \int_x^{\sqrt{V}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|_2)} d\varepsilon \lesssim \sqrt{n}.$$

Suppose $\sqrt{V} \le T$ and

$$x_n^* \lesssim \frac{V}{T}, \quad \text{and} \quad \mathcal{H}_{[\,]}(\sqrt{V}, \mathcal{F}, \|\cdot\|_2) \lesssim \frac{n(x_n^*)^2}{V}.$$

Then, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \hat{\mathbb{E}} f - \mathbb{E} f \ge x_n^* \right) \lesssim \exp\left(-\frac{n(x_n^*)^2}{V + T x_n^*} \right).$$

7.3.4 Proof of Theorem 3

Proof of Theorem 3. By definition of $\hat{\Theta}$, we have

$$\begin{aligned} \mathrm{MAE}(\hat{\Theta}, \Theta) &= \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi\in\mathcal{H}} \mathrm{sgn}\hat{Z}_\pi - \Theta \right| \\ &\le \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi\in\mathcal{H}} \Big(\mathrm{sgn}\hat{Z}_\pi - \mathrm{sgn}(\Theta - \pi) \Big) \right| + \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi\in\mathcal{H}} \mathrm{sgn}(\Theta - \pi) - \Theta \right| \\ &\le \frac{1}{2H+1} \sum_{\pi\in\mathcal{H}} \mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)) + \frac{1}{H}, \end{aligned} \tag{41}$$

where the last line comes from the triangle inequality and the inequality

$$\left| \frac{1}{2H+1} \sum_{\pi\in\mathcal{H}} \mathrm{sgn}(\Theta(\omega) - \pi) - \Theta(\omega) \right| \le \frac{1}{H}, \quad \text{for all } \omega \in [d_1] \times \cdots \times [d_K].$$

Write $n = |\Omega|$. Now it suffices to bound the first term in (41). For any given $t \ge t_n = \frac{d_{\max} r K \log n}{n}$, define the event

$$A = \left\{ \mathrm{MAE}(\mathrm{sgn}\hat{\mathcal{Z}}_\pi, \mathrm{sgn}(\Theta - \pi)) \lesssim t^{\alpha/(2+\alpha)} + \frac{t}{\rho^2(\pi, \mathcal{N})} \text{ for all } \pi \in \mathcal{H} \right\}.$$

We shall prove that under the event A,

$$\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}}\mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi,\mathrm{sgn}(\Theta-\pi))\lesssim t^{\alpha/(\alpha+2)}+\frac{1+|\mathcal{N}|}{H}+Ht. \tag{42}$$

Theorem 2 implies that the sign estimation accuracy depends on the closeness of $\pi\in\mathcal{H}$ to the mass points in $\mathcal{N}$. Therefore, we partition the level set $\pi\in\mathcal{H}$ based on their closeness to $\mathcal{N}$. Specifically, Define $\mathcal{H}_1\overset{\mathrm{def}}{=}\{\pi\in\mathcal{H}\colon \rho(\pi,\mathcal{N})<\frac{1}{H}\}$ and $\mathcal{H}_2=\mathcal{H}\setminus\mathcal{H}_1$. Notice $|\mathcal{H}_1|\le 2|\mathcal{N}|$. We expand the left hand side of (42) by

$$\begin{aligned}&\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}}\mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi,\mathrm{sgn}(\Theta-\pi))\\ =&\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}_1}\mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi,\mathrm{sgn}(\Theta-\pi))+\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}_2}\mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi,\mathrm{sgn}(\Theta-\pi)).\end{aligned} \tag{43}$$

The first term involves only $2|\mathcal{N}|$ many number of sumnmands thus can be bounded by $4|\mathcal{N}|/(2H+1)$. We bound the second term using the explicit forms of $\rho(\pi,\mathcal{N})$ in the sequence $\pi\in\mathcal{H}_2$. Under the event A, we have

$$\begin{aligned}\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}_2}\mathrm{MAE}(\mathrm{sgn}\hat{\mathcal{Z}}_\pi,\mathrm{sgn}(\Theta-\pi))&\lesssim\frac{1}{2H+1}\sum_{\pi\in\mathcal{H}_2}t^{\alpha/(\alpha+2)}+\frac{t}{2H+1}\sum_{\pi\in\mathcal{H}_2}\frac{1}{\rho^2(\pi,\mathcal{N})}\\ &\le t^{\alpha/(\alpha+2)}+\frac{t}{2H+1}\sum_{\pi\in\mathcal{H}_2}\sum_{\pi'\in\mathcal{N}}\frac{1}{|\pi-\pi'|^2}\\ &\le t^{\alpha/(\alpha+2)}+\frac{t}{2H+1}\sum_{\pi'\in\mathcal{N}}\sum_{\pi\in\mathcal{H}_2}\frac{1}{|\pi-\pi'|^2}\\ &\le t^{\alpha/(\alpha+2)}+2CHt,\end{aligned}$$

where the first inequality uses the property of event A, and the last inequality follows from Lemma 3. Combining the bounds for the two terms in (43) completes the proof for conclusion (42); that is

$$\mathbb{P}\left(\mathrm{MAE}(\hat{\Theta},\Theta)\lesssim t^{\alpha/(\alpha+2)}+\frac{1+|\mathcal{N}|}{H}+Ht\right)\ge\mathbb{P}(A). \tag{44}$$

Based on the proof of Theorem 2 and union bound over $\pi\in\mathcal{H}$, we have, for all $t\ge t_n$,

$$\begin{aligned}\mathbb{P}(A)&\ge 1-\sum_{\pi\in\mathcal{H}}\mathbb{P}\left(\mathrm{MAE}(\mathrm{sgn}\hat{\mathcal{Z}}_\pi,\mathrm{sgn}(\Theta-\pi))\gtrsim t^{\alpha/(\alpha+2)}+\frac{t}{\rho(\pi,\mathcal{N})^2}\right)\\ &\gtrsim 1-(2H+1)\exp(-nt)\gtrsim 1-\exp(-nt+\log H).\end{aligned} \tag{45}$$

We choose $t\asymp t_n\log H$ in (45) so that $\log H$ is negligible compared to nt. Finally, combining (44) and (45) with the choice of t yields

$$\mathrm{MAE}(\hat{\Theta},\Theta)\lesssim\left(\frac{d_{\max}rK\log|\Omega|\log H}{|\Omega|}\right)^{\alpha/(\alpha+2)}+\frac{1+|\mathcal{N}|}{H}+\frac{d_{\max}rK\log|\Omega|}{|\Omega|}H\log H,$$

with at least probability $1-\exp(-d_{\max}rK\log|\Omega|\log H)\ge 1-\exp(-d_{\max}rK\log|\Omega|)$.

□

Lemma 3. *Fix $\pi' \in \mathcal{N}$ and a sequence $\Pi = \{-1, \dots, -1/H, 0, 1/H, \dots, 1\}$ with $H \geq 2$. Then,*

$$\sum_{\pi \in \mathcal{H}_2} \frac{1}{|\pi - \pi'|^2} \leq 4H^2.$$

Proof of Lemma 3. Notice that all points $\pi \in \mathcal{H}_2$ satisfy $|\pi - \pi'| \gtrsim \frac{1}{H}$ for all $\pi' \in \mathcal{N}$ by definition and the fact that Δs is negligible compared to $1/H$. We use this fact to compute the sum

$$\begin{aligned}
\sum_{\pi \in \mathcal{H}_2} \frac{1}{|\pi - \pi'|^2} &= \sum_{\frac{h}{H} \in \mathcal{H}_2} \frac{1}{|\frac{h}{H} - \pi'|^2} \\
&\leq 2H^2 \sum_{h=1}^{H} \frac{1}{h^2} \\
&\leq 2H^2 \left\{ 1 + \int_1^2 \frac{1}{x^2} dx + \int_2^3 \frac{1}{x^2} dx + \cdots + \int_{H-1}^{H} \frac{1}{x^2} dx \right\} \\
&= 2H^2 \left(1 + \int_1^H \frac{1}{x^2} dx \right) \leq 4H^2,
\end{aligned}$$

where the third line uses the monotonicity of $\frac{1}{x^2}$ for $x \geq 1$. □

7.3.5 Proof of Theorem 4

Proof of Theorem 4. Write $\bar{\mathcal{Y}} = \mathcal{Y} - \pi$, $\bar{\Theta} = \Theta - \pi$, and $n = |\Omega|$. Here we consider the estimation

$$\hat{\mathcal{Z}}_\pi = \operatorname*{arg\,min}_{\operatorname{rank}(\mathcal{Z}) \leq r} \sum_{\omega \in \Omega} |\bar{\mathcal{Y}}(\omega)| \times F(\mathcal{Z}(\omega) \operatorname{sgn}(\bar{\mathcal{Y}}(\omega)) + \lambda \|\mathcal{Z}\|_F^2, \tag{46}$$

where $\lambda > 0$ is the penalty parameter and F is a large-margin loss satisfying Assumption 2.

The tensor estimation error (11) directly follows from sign tensor estimation error (10) and the proof of Theorem 3. Therefore, it suffices to prove (10). Our proof uses the same techniques used in the proof of Theorem 2. We summarize only the key difference.

Fix $\pi \notin \mathcal{N}$. For notational simplicity, we suppress the subscript π and write $\hat{\mathcal{Z}}$ in place of $\hat{\mathcal{Z}}_\pi$. Denote $n = |\Omega|$ and $\rho = \rho(\pi, \mathcal{N})$. Define $\ell_{\omega,F}(\mathcal{Z}) = |\bar{\mathcal{Y}}(\omega)| \times F(\mathcal{Z}(\omega) \operatorname{sgn}(\bar{\mathcal{Y}}(\omega))$ and $\ell_{\omega,F'}(\mathcal{Z}) = |\bar{\mathcal{Y}}(\omega)| \times F'(\mathcal{Z}(\omega) \operatorname{sgn}(\bar{\mathcal{Y}}(\omega))$ where F' is T-truncated version of F such that $F'(x) = \min(F(x), T)$ with $T = \max(2, J^2)$. We focus on the following two empirical processes induced by function $f_{\mathcal{Z}} \in \mathcal{F}_{\mathcal{T}}$ where $\mathcal{T} = \{\mathcal{Z} \colon \operatorname{rank}(\mathcal{Z}) \leq r\}$,

$$\frac{1}{n} \sum_{i=1}^{n} \underbrace{\left[\ell_{i,F}(\mathcal{Z}, \bar{\mathcal{Y}}) - \ell_{i,F}(\bar{\Theta}, \bar{\mathcal{Y}}) \right]}_{\stackrel{\text{def}}{=} \Delta_{i,F}(f_{\mathcal{Z}}, \bar{\Theta})}, \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left[\ell_{i,F'}(\mathcal{Z}, \bar{\mathcal{Y}}) - \ell_{i,F'}(\bar{\Theta}, \bar{\mathcal{Y}}) \right]}_{\stackrel{\text{def}}{=} \Delta_{i,F'}(f_{\mathcal{Z}}, \bar{\Theta})}.$$

Note that there is an one-to-one correspondence between sets $\mathcal{F}_{\mathcal{T}}$ and $\mathcal{T}$.

By definition of $\hat{\mathcal{Z}}$ in (46), we have

$$\frac{1}{n} \sum_{i=1}^{n} \Delta_{i,F}(f_{\hat{\mathcal{Z}}}, \mathcal{Z}^{(n)}) \leq \lambda J^2 - \lambda \|\hat{\mathcal{Z}}\|_F^2,$$

where $\mathcal{Z}^{(n)}$ is a sequence of function in Assumption 2(a). Let L_n denote the desired convergence rate to seek. Then, we have the following inclusion of probability events,

$$
\begin{aligned}
&\left\{(\omega, \mathcal{Y}_\omega)\colon \mathrm{Risk}_{F'}(\hat{\mathcal{Z}}) - \mathrm{Risk}_{F'}(\bar{\Theta}) \geq 2L_n\right\} \\
\subset &\left\{(\omega, \mathcal{Y}_\omega)\colon \exists \mathcal{Z} \text{ s.t. } \mathrm{rank}(\mathcal{Z}) \leq r, \mathrm{Risk}_{F'}(\mathcal{Z}) - \mathrm{Risk}_{F'}(\bar{\Theta}) \geq 2L_n, \right. \\
&\qquad\qquad \left. \text{and } -\frac{1}{n}\sum_{i=1}^n \Delta_{i,F}(f_{\mathcal{Z}}, \mathcal{Z}^{(n)}) + \lambda J^2 - \lambda \|\hat{\mathcal{Z}}\|_F^2 \geq 0\right\} \\
\overset{(*)}{\subset} &\left\{(\omega, \mathcal{Y}_\omega)\colon \exists \mathcal{Z} \text{ s.t. } \mathrm{rank}(\mathcal{Z}) \leq r, \mathrm{Risk}_{F'}(\mathcal{Z}) - \mathrm{Risk}_{F'}(\bar{\Theta}) \geq 2L_n, \right. \\
&\qquad\qquad \left. \text{and } -\frac{1}{n}\sum_{i=1}^n \Delta_{i,F'}(f_{\mathcal{Z}}, \mathcal{Z}^{(n)}) + \lambda J^2 - \lambda \|\hat{\mathcal{Z}}\|_F^2 \geq 0\right\} \\
\subset &\left\{(\omega, \mathcal{Y}_\omega)\colon \sup_{\substack{\mathrm{rank}(\mathcal{Z}) \leq r \\ \mathrm{Risk}_{F'}(\mathcal{Z}) - \mathrm{Risk}_{F'}(\bar{\Theta}) \geq 2L_n}} -\frac{1}{n}\sum_{i=1}^n \Delta_{i,F'}(f_{\mathcal{Z}}, \mathcal{Z}^{(n)}) + \lambda J^2 - \lambda \|\hat{\mathcal{Z}}\|_F^2 \geq 0\right\} \\
\subset &\bigcup_{\ell_1,\ell_2=1}^{\infty} \left\{(\omega, \mathcal{Y}_\omega)\colon \sup_{\mathcal{Z} \in A_{\ell_1,\ell_2}} -\frac{1}{n}\sum_{i=1}^n \Delta_{i,F'}(f_{\mathcal{Z}}, \mathcal{Z}^{(n)}) + \lambda J^2 - \lambda \|\hat{\mathcal{Z}}\|_F^2 \geq 0\right\}, \qquad (47)
\end{aligned}
$$

where $(*)$ comes from the fact

$$
\ell_{\omega,F'}(\mathcal{Z}, \bar{\mathcal{Y}}) \leq \ell_{\omega,F}(\mathcal{Z}, \bar{\mathcal{Y}}) \text{ for all } \mathcal{Z}, \quad \text{and } \ell_{\omega,F'}(\mathcal{Z}^{(n)}, \bar{\mathcal{Y}}) = \ell_{\omega,F}(\mathcal{Z}^{(n)}, \bar{\mathcal{Y}}),
$$

because the truncation constant $T = \max(2, J^2) \geq \max(2, \sup_n \|\mathcal{Z}^{(n)}\|_F^2)$. In the last line of (47), we have partitioned $\{\mathcal{Z}\colon \mathrm{rank}(\mathcal{Z}) \leq r \text{ and } \mathrm{Risk}_{F'}(\mathcal{Z}) - \mathrm{Risk}_{F'}(\bar{\Theta}) \geq 2L_n\}$ into union of A_{ℓ_1,ℓ_2} with

$$
\begin{aligned}
A_{\ell_1,\ell_2} = &\left\{\mathcal{Z}\colon \mathrm{rank}(\mathcal{Z}) \leq r, (\ell_1+1)L_n \leq \mathrm{Risk}_{F'}(\mathcal{Z}) - \mathrm{Risk}_{F'}(\bar{\Theta}) < (\ell_1+2)L_n, \right. \\
&\qquad \left. \text{and } (\ell_2 - 1)J^2 \leq \|\mathcal{Z}\|_F^2 < \ell_2 J^2\right\},
\end{aligned}
$$

for $\ell_1, \ell_2 = 1, 2, \ldots$.

Let Γ denote the target probability for the first line in (47). For each A_{ℓ_1,ℓ_2}, we consider the centered empirical process,

$$
v_n(f_{\mathcal{Z}}) := -\frac{1}{n}\sum_{i=1}^n \left(\Delta_{i,F'}(f_{\mathcal{Z}}, \mathcal{Z}^{(n)}) - \mathbb{E}\Delta_{i,F'}(f_{\mathcal{Z}}, \mathcal{Z}^{(n)})\right). \qquad (48)
$$

Notice that

$$
\begin{aligned}
\mathbb{E}\Delta_{i,F'}(f_{\mathcal{Z}}, \mathcal{Z}^{(n)}) &= \mathrm{Risk}_{F'}(\mathcal{Z}) - \mathrm{Risk}_{F'}(\bar{\Theta}) + \mathrm{Risk}_{F'}(\bar{\Theta}) - \mathrm{Risk}_{F'}(\mathcal{Z}^{(n)}) \\
&\geq (\ell_1+1)L_n - a_n \\
&\geq \ell_1 L_n,
\end{aligned}
$$

where the first inequality is from the fact that $\mathcal{Z} \in A_{\ell_1,\ell_2}$ and Assumption 2(a), and the last inequality uses the condition that $a_n \lesssim L_n$.

Combining (47), (48) and the union bound yields

$$\Gamma \leq \sum_{\ell_1,\ell_2=1}^{\infty} \mathbb{P}\left\{ \sup_{\mathcal{Z}\in A_{\ell_1,\ell_2}} v_n(f_{\mathcal{Z}}) \geq \ell_1 L_n + \lambda(\ell_2-2)J^2 =: M(\ell_1,\ell_2) \right\}. \tag{49}$$

Similar to the proof of Lemma 1 and Lemma 2 with T-truncated hinge loss in Lee et al. (2021), the variance of empirical process is bounded by

$$\begin{aligned}\sup_{\mathcal{Z}\in A_{\ell_1,\ell_2}} \mathrm{Var}\Delta_{i,F'}(f_{\mathcal{Z}},\bar{\Theta}) &\lesssim \sup_{\mathcal{Z}\in A_{\ell_1,\ell_2}} \left(\left[\mathbb{E}\Delta_{i,F'}(f_{\mathcal{Z}},\bar{\Theta})\right]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho}\mathbb{E}\Delta_{i,F'}(f_{\mathcal{Z}},\bar{\Theta}) \right) + \Delta s \\ &\lesssim M(\ell_1,\ell_2)^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho}M(\ell_1,\ell_2) + \Delta s =: V(\ell_1,\ell_2).\end{aligned}$$

To apply Theorem 5, we choose the pair (L_n,λ) satisfying

$$\sup_{\ell_1,\ell_2\geq 1} \frac{1}{x}\int_x^{\sqrt{x^{\alpha/(\alpha+1)}+x/\rho+\Delta s}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon,\mathcal{F}_{\mathcal{T}}(\ell_2),\|\cdot\|_2)}d\varepsilon \lesssim n^{1/2}, \tag{50}$$

where $x = \ell_1 L_n + \lambda(\ell_2-2)J^2$ and $\mathcal{F}_{\mathcal{T}}(\ell_2) := \{f_{\mathcal{Z}} \colon \mathrm{rank}(\mathcal{Z}) \leq r, \|\mathcal{Z}\|_F^2 \leq \ell_2 J^2\}$. Similar to the proof of Lemma 2, we solve the pair (L_n,λ) satisfying (50) as

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + \frac{t_n}{\rho}, \quad \text{and} \quad \lambda = \frac{L_n}{2J^2}, \tag{51}$$

where $t_n = \frac{d_{\max} r K \log n}{n}$. With the choice (51), we bound the right-hand side of (49) based on Theorem 5,

$$\begin{aligned}\Gamma &\lesssim \sum_{\ell_1,\ell_2=1}^{\infty} \exp\left(-\frac{nM^2(\ell_1,\ell_2)}{V(\ell_1,\ell_2)+2M(\ell_1,\ell_2)}\right) \\ &\lesssim \sum_{\ell_1,\ell_2=1}^{\infty} \exp(-\rho n M(\ell_1,\ell_2)) \\ &\leq \left(\frac{e^{-n\rho L_n}}{1-e^{-n\rho L_n}}\right)\left(\frac{e^{n\rho\lambda J^2}}{1-e^{-n\rho\lambda J^2}}\right) \\ &\lesssim e^{-n\rho L_n} \leq e^{-nt_n},\end{aligned}$$

where the last line uses the fact that $2\rho\lambda J^2 = \rho L_n \gtrsim t_n \gtrsim n^{-1}$ from (51). The proof is then completed by (37). □

7.4 Extension of Theorems 2-3 to unbounded observation with sub-Gaussian noise

Consider the signal plus noise model

$$\mathcal{Y} = \Theta + \mathcal{E},$$

where $\mathcal{E}$ consists of zero-mean, independent noise entries, and $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$ is an α-smooth tensor. Theoretical results in Section 4 of the main paper are based on bounded observation $\|\mathcal{Y}\|_\infty \leq 1$. We extend the results to unbounded observation with the following assumption.

Assumption 3 (Sub-Gaussian noise)**.**

1. There exists a constant $\beta > 0$, independent of tensor dimension, such that $\|\Theta\|_\infty \le \beta$. Without loss of generality, we set $\beta = 1$.
2. The noise entries $\mathcal{E}(\omega)$ are independent zero-mean sub-Gaussian random variables with variance proxy $\sigma^2 > 0$; i.e, $\mathbb{P}(|\mathcal{E}(\omega)| \ge B) \le 2e^{-B^2/2\sigma^2}$ for all $B > 0$.

We say that an event A occurs "with high probability" if $\mathbb{P}(A)$ tends to 1 as the tensor dimension $d_{\min} = \min_k d_k \to \infty$. The following result show that the sub-Gaussian noise incurs an additional $\log|\Omega|$ factor compared to the bounded case.

Theorem 6 (Extension to sub-Gaussian noise)**.** *Consider the same condition of Theorem 2. Suppose that Assumption 3 holds. With high probability over training data $\mathcal{Y}_\Omega$, we have:*

(a) (Sign matrix estimation). For all $\pi \notin \mathcal{N}$,

$$\mathrm{MAE}(\mathrm{sgn}(\hat{\mathcal{Z}}_\pi), \mathrm{sgn}(\Theta - \pi)) \lesssim t_d^{\frac{\alpha}{\alpha+2}} + \frac{t_d}{\rho^2(\pi, \mathcal{N})}, \text{ where } t_d := \frac{r\sigma^2 d_{\max} \log d_{\max} \log|\Omega|}{|\Omega|}.$$

(b) For all resolution parameter $H \in \mathbb{N}_+$,

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim \left(\sigma^2 t_d \log d_{\max} \log H\right)^{\alpha/(\alpha+2)} + \frac{1 + |\mathcal{N}|}{H} + H(\sigma^2 t_d \log d_{\max} \log H). \tag{52}$$

In particular, setting $H \asymp \left(\frac{1+|\mathcal{N}|}{\sigma^2 t_d \log d_{\max}}\right)^{1/2}$ yields the tightest upper bound in (52)*.*

Proof of Theorem 6. By setting $s = K\log(d_{\max})$ in Lemma 4, we have

$$\mathbb{P}(\|\mathcal{E}\|_\infty \ge \sqrt{4\sigma^2 K \log d_{\max}}) \le 2d_{\max}^{-K}.$$

We divide the sample space into two exclusive events:

- Event I: $\|\mathcal{E}\|_\infty \ge \sqrt{4\sigma^2 K \log d_{\max}}$;
- Event II: $\|\mathcal{E}\|_\infty < \sqrt{4\sigma^2 K \log d_{\max}}$.

Because the Event I occurs with probability tending to zero, we restrict ourselves to the Event II only, by following the proof of Theorem 2. We summarize the key difference compared to Section 7.3. We expand the variance by

$$\begin{aligned}
\mathrm{Var}\left[\ell_\omega\left(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega\right) - \ell_\omega\left(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega\right)\right] &\le \mathbb{E}|\ell_\omega(\mathcal{Z}(\omega), \bar{\mathcal{Y}}(\omega)) - \ell_\omega(\bar{\Theta}(\omega), \bar{\mathcal{Y}}(\omega))|^2 \\
&= \mathbb{E}|\bar{\mathcal{Y}}(\omega) - \bar{\Theta}(\omega) + \bar{\Theta}(\omega)|^2 |\mathrm{sgn}\mathcal{Z}(\omega) - \mathrm{sgn}\bar{\Theta}(\omega)| \\
&\le 2\left(4\sigma^2 K \log d_{\max} + 2\right)\mathbb{E}|\mathrm{sgn}\mathcal{Z} - \mathrm{sgn}\bar{\Theta}| \\
&\lesssim (\sigma^2 K \log d_{\max})\mathrm{MAE}(\mathrm{sgn}\mathcal{Z}, \mathrm{sgn}\bar{\Theta}),
\end{aligned} \tag{53}$$

where the third line uses the facts $\|\bar{\Theta}\|_\infty \le 2$ and $\|\bar{\mathcal{Y}} - \bar{\Theta}\|_\infty^2 = \|\mathcal{E}\|_\infty^2 < 4\sigma^2 K \log d_{\max}$ within the Event II; the last line comes from the definition of MAE and the asymptotic $\sigma^2 \log d_{\max} \gg 1$ provided that $\sigma > 0$ with $d_{\max}$ sufficiently large.

Based on (53), the α-smoothness of Θ implies that for all measurable functions $f_{\mathcal{Z}}$, we have

$$\mathrm{Var}\Delta_i(f_{\mathcal{Z}}, \bar{\Theta}) \lesssim \left(\sigma^2 K \log d_{\max}\right)\left\{\left[\mathbb{E}\Delta_i(f_{\mathcal{Z}}, \bar{\Theta})\right]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho}\mathbb{E}\Delta_i(f_{\mathcal{Z}}, \bar{\Theta}) + \Delta s\right\}. \tag{54}$$

Based on the proof of Theorem 2, the empirical process with variance-to-mean relationship (54) gives that

$$\mathbb{P}\left(\mathrm{Risk}(\hat{\mathcal{Z}}) - \mathrm{Risk}(\bar{\Theta}) \geq L_n\right) \lesssim \exp(-nt_n), \tag{55}$$

where the convergence rate L_n is obtained by the same way in the proof of Lemma 2,

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho} t_n, \quad \text{with } t_n = \frac{r\sigma^2 d_{\max} \log d_{\max} \log n}{n}, \tag{56}$$

where constants (possibly depending on K) have been absorbed into the $\asymp$ relationship. Combining (55) and (56), we obtain that, with high probability,

$$\mathrm{Risk}(\hat{\mathcal{Z}}) - \mathrm{Risk}(\bar{\Theta}) \lesssim \left(\frac{r\sigma^2 d_{\max} \log d_{\max} \log |\Omega|}{|\Omega|}\right)^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi, \mathcal{N})} \left(\frac{r\sigma^2 d_{\max} \log d_{\max} \log |\Omega|}{|\Omega|}\right), \tag{57}$$

Therefore, combining (57) and (37) completes the proof. The tensor estimation error follows readily from the proof of Theorem 3 and Theorem 6. □

Lemma 4 (sub-Gaussian maximum)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian zero-mean random variables with variance proxy σ^2. Then, for any $s > 0$,*

$$\mathbb{P}\left\{\max_{1 \leq i \leq n} |X_i| \geq \sqrt{2\sigma^2(\log n + s)}\right\} \leq 2e^{-s}.$$

Proof of Lemma 4. The conclusion follows from

$$\mathbb{P}[\max_{1 \leq i \leq n} |X_i| \geq u] \leq \sum_{i=1}^{n} \mathbb{P}[X_i \geq u] \leq 2ne^{-\frac{u^2}{2\sigma^2}} = 2e^{-s},$$

where we set $u = \sqrt{2\sigma^2(\log n + s)}$. □

8 Conclusion

We have developed a tensor estimation method that addresses both low- and high-rankness. Our sign series representation leads to nonparametric guarantees previously impossible. We hope the work opens up new inquiry that allows more researchers to contribute to this field.

Acknowledgements

This research is supported in part by NSF grant DMS-1915978, DMS-2023239 and Wisconsin Alumni Research Foundation.

References

Alon, N., Moran, S., and Yehudayoff, A. (2016). Sign rank versus VC dimension. In *Conference on Learning Theory*, pages 47–80.

Alquier, P., Cottet, V., Lecué, G., et al. (2019). Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *Annals of Statistics*, 47(4):2117–2144.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832.

Anandkumar, A., Ge, R., and Janzamin, M. (2017). Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research*, 18(1):752–791.

Balabdaoui, F., Durot, C., and Jankowski, H. (2019). Least squares estimation in the monotone single index model. *Bernoulli*, 25(4B):3276–3310.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Cai, C., Li, G., Poor, H. V., and Chen, Y. (2019). Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pages 1863–1874.

Chan, S. and Airoldi, E. (2014). A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216.

Chi, E. C., Gaines, B. J., Sun, W. W., Zhou, H., and Yang, J. (2020). Provable convex co-clustering of tensors. *Journal of Machine Learning Research*, 21(214):1–58.

Cohn, H. and Umans, C. (2013). Fast matrix multiplication using coherent configurations. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1074–1087.

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

De Wolf, R. (2003). Nondeterministic quantum query and communication complexities. *SIAM Journal on Computing*, 32(3):681–699.

Fan, J. and Udell, M. (2019). Online high rank matrix completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8690–8698.

Ganti, R., Rao, N., Balzano, L., Willett, R., and Nowak, R. (2017). On learning high dimensional structured single index models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 1898–1904.

Ganti, R. S., Balzano, L., and Willett, R. (2015). Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, pages 1873–1881.

Genzel, M. and Stollenwerk, A. (2020). Robust 1-bit compressed sensing via hinge loss minimization. *Information and Inference: A Journal of the IMA*, 9(2):361–422.

Ghadermarzy, N., Plan, Y., and Yilmaz, O. (2018). Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing*, 67(1):29–40.

Ghadermarzy, N., Plan, Y., and Yilmaz, Ö. (2019). Near-optimal sample complexity for convex tensor completion. *Information and Inference: A Journal of the IMA*, 8(3):577–619.

Globerson, A., Chechik, G., Pereira, F., and Tishby, N. (2007). Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, 8:2265–2295.

Han, R., Willett, R., and Zhang, A. (2020). An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, L., Lu, C.-T., Ma, G., Wang, S., Shen, L., Philip, S. Y., and Ragin, A. B. (2017). Kernelized support tensor machines. In *International Conference on Machine Learning*, pages 1442–1451. PMLR.

Hillar, C. J. and Lim, L.-H. (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):45.

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

Hong, D., Kolda, T. G., and Duersch, J. A. (2020). Generalized canonical polyadic tensor decomposition. *SIAM Review*, 62(1):133–163.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094.

Jain, P. and Oh, S. (2014). Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, volume 27, pages 1431–1439.

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

Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

Lee, C., Li, L., Zhang, H. H., and Wang, M. (2021). Nonparametric trace regression in high dimensions via sign series representation. *arXiv preprint arXiv:2105.01783*.

Lee, C. and Wang, M. (2020). Tensor denoising and completion based on ordinal observations. In *International Conference on Machine Learning*, pages 5778–5788.

Li, Y., Liu, Y., Li, J., Qin, W., Li, K., Yu, C., and Jiang, T. (2009). Brain anatomical network and intelligence. *PLoS Comput Biol*, 5(5):e1000395.

Montanari, A. and Sun, N. (2018). Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics*, 71(11):2381–2425.

Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, pages 73–81.

Ongie, G., Willett, R., Nowak, R. D., and Balzano, L. (2017). Algebraic variety models for high-rank matrix completion. In *International Conference on Machine Learning*, pages 2691–2700.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 56(4):931–954.

Scott, C. (2011). Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In *ICML*.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, 22:580–615.

Wang, L., Durante, D., Jung, R. E., and Dunson, D. B. (2017). Bayesian network–response regression. *Bioinformatics*, 33(12):1859–1866.

Wang, M. and Li, L. (2020). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *Journal of Machine Learning Research*, 21(154):1–38.

Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, pages 713–723.

Yuan, M. and Zhang, C.-H. (2016). On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068.

Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311 – 7338.