Beyond the Signs: Nonparametric Tensor Completion via Sign Series

Chanwoo Lee
University of Wisconsin – Madison
chanwoo.lee@wisc.edu

Miaoyan Wang
University of Wisconsin – Madison
miaoyan.wang@wisc.edu

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, certain hypergraphon models—as special cases. We show that the sign tensor series is theoretically characterized, and computationally estimable, via classification tasks with carefully-specified weights. Excess risk bounds, estimation error rates, 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?

1.1 Inadequacies of low-rank models

The signal plus noise model (3) 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 signal recovery, 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 CP rank$(\mathcal{Z}) = 3$ (formal definition is deferred to 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 Appendix) 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.
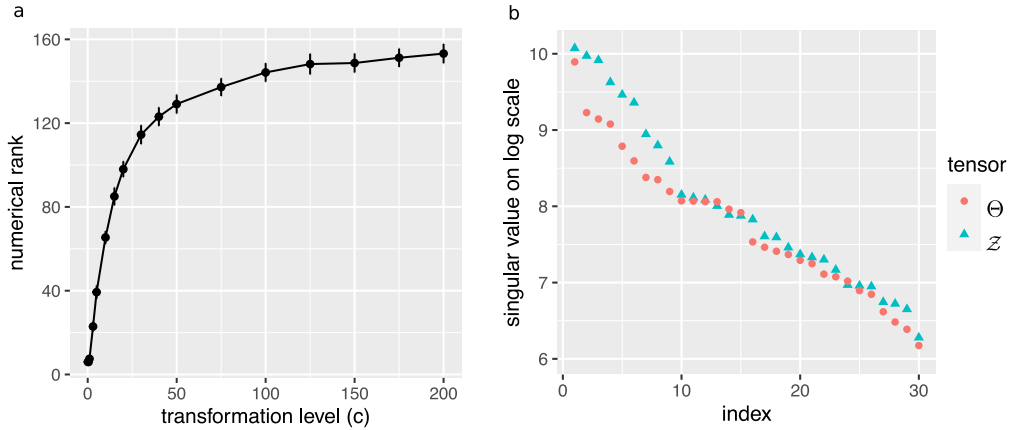

Figure 1: (a) Numerical rank of Θ versus c in the first example. (b) Top $d = 30$ tensor singular values in the second example.

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 $\mathcal{Z}(i, j, k) = \frac{1}{d} \max(i, j, k)$ for $i, j, k \in \{1, \ldots, d\}$. The matrix analogy of Θ was studied by Chan and Airoldi (2014) in graphon analysis. 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. New methods that allow flexible tensor modeling have yet to be developed.

1.2 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
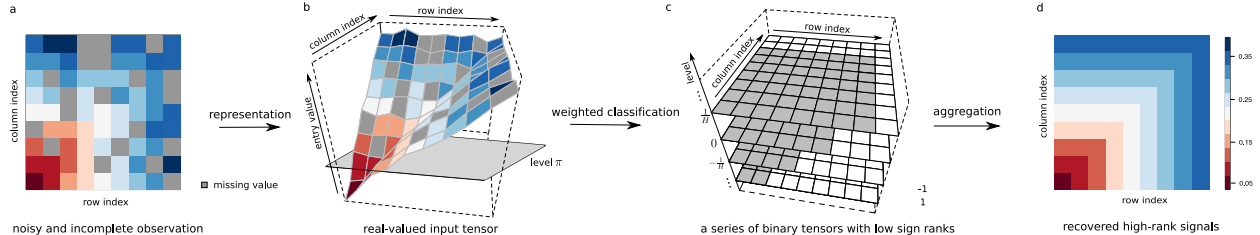
Figure 2: Illustration of our method. For visualization purpose, we plot an order-2 tensor (a.k.a. matrix); similar procedure applies to higher-order tensors. (a): a noisy and incomplete tensor input. (b) and (c): main steps of estimating sign tensor series $\text{sgn}(\Theta - \pi)$ for $\pi \in \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$. (d) estimated signal $\hat{\Theta}$. The depicted signal is a full-rank matrix based on Example 5 in Section 3.

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.

From computational perspective, optimizations regarding tensors are in general NP-hard. Fortunately, tensors sought in applications are specially-structured, for which a number of efficient algorithms are available (Ghadermarzy et al., 2018; Wang and Li, 2020; Han et al., 2020). Our high-rank tensor estimate is provably reductable to a series of classifications, and its divide-and-conquer nature facilitates efficient computation. The ability 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, the problem for high-rank tensors is more challenging, because the tensor rank often exceeds the dimension when order $K \geq 3$ (Anandkumar et al., 2017). This is in sharp contrast to matrices. 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.

1.3 Notation

We use $\text{sgn}(\cdot) \colon \mathbb{R} \to \{-1, 1\}$ to denote the sign function, where $\text{sgn}(y) = 1$ if $y \geq 0$ and -1 otherwise. We allow univariate functions, such as $\text{sgn}(\cdot)$ and general $f \colon \mathbb{R} \to \mathbb{R}$, to be applied to tensors in an element-wise manner. We denote $a_n \lesssim b_n$ if $\lim_{n \to \infty} a_n/b_n \leq c$ for some constant $c \geq 0$. We use the shorthand $[n]$ to denote the n-set $\{1, \ldots, n\}$ for $n \in \mathbb{N}_+$. Let $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K $(d_1, \ldots, d_K)$-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 E is said to occur "with very high probability" if $\mathbb{P}(E)$ tends to 1 faster than any polynomial of tensor dimension $d := \min_k d_k \to \infty$. The CP decomposition (Hitchcock, 1927) is defined by

$$\Theta = \sum_{s=1}^{r} \lambda_s \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}, \tag{2}$$

3

where $\lambda_1 \geq \cdots \geq \lambda_r > 0$ are tensor singular values, $\boldsymbol{a}_s^{(k)} \in \mathbb{R}^{d_k}$ are norm-1 tensor singular vectors, and $\otimes$ denotes the outer product of vectors. The minimal $r \in \mathbb{N}_+$ for which (2) holds is called the tensor rank, denoted $\text{rank}(\Theta)$.

2 Model and proposal overview

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

$$\mathcal{Y} = \Theta + \mathcal{E}, \tag{3}$$

where $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is an unknown signal tensor of interest, and $\mathcal{E}$ is a noise tensor consisting of mean-zero, 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. Assume that $\mathcal{Y}(\omega)$ takes value in a bounded interval $[-A, A]$; without loss of generality, we set $A = 1$ throughout the paper.

Our observation is an incomplete data tensor from (3), 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 $\sum_{\omega \in [d_1] \times \cdots \times [d_K]} p_\omega = 1$. Assume that the entries ω in Ω are i.i.d. draws with replacement from the full index set using distribution Π. The sampling rule is denoted as $\omega \sim \Pi$.

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 $\text{sgn}(\Theta - \pi)$ 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 $\text{sgn}(\Theta - \pi) = \text{sgn}(\mathcal{Z} - f^{-1}(\pi))$. The signal tensor in the second example has the same sign pattern as a rank-2 tensor, since $\text{sgn}(\Theta - \pi) = \text{sgn}(\max(i, j, k) - d(e^\pi - 1))$ (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 dichotomize the data tensor into a series of sign tensors $\text{sgn}(\mathcal{Y}_\Omega - \pi)$ for $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$. Then, we estimate the sign signals $\text{sgn}(\Theta - \pi)$ by performing classification

$$\hat{\mathcal{Z}}_\pi = \operatorname*{arg\,min}_{\text{low rank tensor } \mathcal{Z}} \text{Weighted-Loss}(\text{sgn}(\mathcal{Z}), \text{sgn}(\mathcal{Y}_\Omega - \pi)),$$

where Weighted-Loss$(\cdot, \cdot)$ denotes a carefully-designed classification objective function which will be described in later sections. Our final proposed tensor estimate takes the form

$$\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \text{sgn}(\hat{\mathcal{Z}}_\pi).$$

Our approach is built on the nonparametric sign representation of signal tensors. The estimate $\hat{\Theta}$ is essentially learned from dichotomized tensor series $\{\text{sgn}(\mathcal{Y}_\Omega - \pi) : \pi \in \mathcal{H}\}$ with proper weights. 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 that were previously impossible. The method enjoys statistical effectiveness and computational efficiency.

3   Statistical properties of sign representable tensors

This section develops sign representable tensor models for Θ in (3). We characterize the algebraic and statistical properties of sign tensor series, which serves the theoretical foundation for our method.

3.1   Sign-rank and sign tensor series

Let Θ be the tensor of interest, and $\text{sgn}(\Theta)$ 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.,

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

The sign-rank is also called *support rank* (Cohn and Umans, 2013), *minimal rank* (Alon et al., 2016), and *nondeterministic rank* (De Wolf, 2003). Earlier work defines sign-rank for binary-valued tensors; 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, $\text{srank}(\Theta) = \text{srank}(\text{sgn}\Theta)$.

Like most tensor problems (Hillar and Lim, 2013), determining the sign-rank for a general tensor is NP hard (Alon et al., 2016). Fortunately, tensors arisen in applications often possess special structures that facilitate analysis. By definition, the sign-rank is upper bounded by the tensor rank. More generally, we have the following upper bounds.

Proposition 1 (Upper bounds of the sign-rank)**.** *For any strictly monotonic function $g \colon \mathbb{R} \to \mathbb{R}$ with $g(0) = 0$,*
$$\text{srank}(\Theta) \leq \text{rank}(g(\Theta)).$$

Conversely, the sign-rank can be much smaller than the tensor rank, as we have shown in the examples of Section 1.

Proposition 2 (Broadness)**.** *For every order $K \geq 2$ and dimension d, there exist tensors $\Theta \in \mathbb{R}^{d \times \cdots \times d}$ such that $\text{rank}(\Theta) \geq d$ but $\text{srank}(\Theta - \pi) \leq 2$ for all $\pi \in \mathbb{R}$.*

We provide several 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. Propositions 1 and 2 together demonstrate the strict broadness of low sign-rank family over the usual low-rank family.

We now introduce a tensor family, which we coin as "sign representable tensors", for the signal model in (3).

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 $\{\text{sgn}(\Theta - \pi) \colon \pi \in$

$[-1, 1]\}$ is called the sign tensor series. We use $\mathscr{P}_{\text{sgn}}(r) = \{\Theta \colon \text{srank}(\Theta - \pi) \leq r \text{ for all } \pi \in [-1, 1]\}$ to denote the r-sign representable tensor family.

We 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 certain hypergraphon models.

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., $\Theta \in \mathscr{P}_{\text{sgn}}(r + 1)$ (the constant 1 is due to rank$(\Theta - \pi) \leq r + 1$). Similar results hold for Tucker low-rank tensors $\Theta \in \mathscr{P}_{\text{sgn}}(r + 1)$, 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 checkerbord 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 $\Theta \in \mathscr{P}_{\text{sgn}}(r)$.

Example 3 (Generalized linear models (GLMs))**.** Let $\mathcal{Y}$ be a binary tensor from a logistic model (Wang and Li, 2020) with mean $\Theta = \text{logit}(\mathcal{Z})$, where $\mathcal{Z}$ is a latent low-rank tensor. Notice that Θ itself may be high-rank (see Section 1). 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). We here extend the model to higher-order tensors Θ. 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 (Min/Max hypergraphon)**.** Graphon is a popular nonparametric model for networks (Chan and Airoldi, 2014; Xu, 2018). Here we revisit the model introduced in Section 1 for generality. Let Θ be an order-K tensor generated from the hypergraphon $\Theta(i_1, \ldots, i_K) = \log(1 + \max_k x_{i_k}^{(k)})$, where $x_{i_k}^{(k)}$ are given number 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).

The results extend to general min/max hypergraphons. Let $g(\cdot)$ be a continuous univariate function with at most $r \geq 1$ distinct real roots in the equation $g(z) = \pi$; this property holds, e.g., when $g(z)$ is a polynomial of degree r. Then, the tensor Θ generated from $\Theta(i_1, \ldots, i_K) = g(\max_k x_{i_k}^{(k)})$ belongs to $\mathscr{P}_{\text{sgn}}(2r)$ (see Appendix). Same conclusion holds if the maximum in $g(\cdot)$ is replaced by the minimum.

3.2 Statistical characterization of sign tensors via weighted classification

Accurate estimation of a sign representable tensor depends on the behavior of sign tensor series, $\text{sgn}(\Theta - \pi)$. In this section, we show that sign tensors are completely 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 (3) and the sampling distribution $\omega \sim \Pi$. Note that the form of $\text{Risk}(\cdot)$ implicitly depends on π; we suppress π when no confusion arises.

Proposition 3 (Global optimum of weighted risk). *Suppose the data $\mathcal{Y}_\Omega$ is generated from model (3) with $\Theta \in \mathscr{P}_{\text{sgn}}(r)$. Then, for all $\bar{\Theta}$ that are sign equivalent to $\text{sgn}(\Theta - \pi)$,*

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

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 turns out to play a key role in the accuracy.

Some additional notation is needed. We use $\mathcal{N} = \{\pi \colon \mathbb{P}_{\omega \sim \Pi}(\Theta(\omega) = \pi) \neq 0\}$ to denote the set of mass points of Θ under Π. Assume there exists a constant $C > 0$, independent of tensor dimension, such that $|\mathcal{N}| \leq C$. Note that both Π and Θ implicitly depend on the tensor dimension. Our assumptions are imposed to $\Pi = \Pi(d)$ and $\Theta = \Theta(d)$ in the high-dimensional regime uniformly where $d := \min_k d_k \to \infty$.

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

$$\sup_{0 \leq t < \rho(\pi, \mathcal{N})} \frac{\mathbb{P}_{\omega \sim \Pi}[|\Theta(\omega) - \pi| \leq t]}{t^\alpha} \leq c, \tag{6}$$

where $\rho(\pi, \mathcal{N}) := \min_{\pi' \in \mathcal{N}} |\pi - \pi'|$ denotes the 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 set $\{\omega : |\Theta(\omega) - \pi| \leq t\}$ has zero measure, implying almost no entries of which $\Theta(\omega)$ is around the level π. We call a tensor Θ is α-globally smooth, if (6) holds with a global constant $c > 0$ for all $\pi \in [-1, 1]$ except for a finite number of levels.

The smoothness index α quantifies the intrinsic hardness of recovering $\text{sgn}(\Theta - \pi)$ 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 the cumulative distribution function (CDF) $G(\pi) := \mathbb{P}_{\omega \sim \Pi}[\Theta(\omega) \leq \pi]$ remains flat around π. A small value of $\alpha < 1$ indicates the nonexistent (infinite) density at level π, or equivalently, when the $G(\pi)$ jumps at π. A typical case is $\alpha = 1$ when the $G(\pi)$ has finite non-zero derivative in the vicinity of π. Table 1 illustrates the $G(\pi)$ for various models of Θ (see Section 5 Simulation for details).

We now reach the main theorem in this section. For two tensors Θ_1, Θ_2, define the mean absolute error (MAE)

$$\text{MAE}(\Theta_1, \Theta_2) \overset{\text{def}}{=} \mathbb{E}_{\omega \sim \Pi} |\Theta_1(\omega) - \Theta_2(\omega)|.$$

Theorem 1 (Identifiability). *Under Assumption 1, for all tensors $\bar{\Theta} \simeq \text{sgn}(\Theta - \pi)$ and tensors $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$,*

$$\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) \leq C(\pi) \left[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) \right]^{\alpha/(\alpha+1)},$$

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

The result establishes the recovery stability of sign tensors $\text{sgn}(\Theta - \pi)$ using optimization with population risk (5). The bound immediately shows the uniqueness of the optimizer for $\text{Risk}(\cdot)$ up to a zero-measure set under Π, as long as $\alpha \neq 0$. 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 for now. The tensor block model is ∞-globally smooth. This is because the set $\mathcal{N}$, which consists of distinct block means in Θ, has finitely many elements. Furthermore, we have $\alpha = \infty$ for all $\pi \notin \mathcal{N}$, since the numerator in (6) is zero for all such π. The min/max hypergaphon model with a r-degree polynomial function is 1-globally smooth because $\alpha = 1$ for all π in the function range except at most $(r - 1)$ many stationary points.

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 algorithm proposed in Section 2 (Figure 2) in details. We provide the estimation error bound and address the empirical implementation of the algorithm.

4.1 Estimation error and sample complexity

Given a noisy incomplete tensor observation $\mathcal{Y}_\Omega$ from model (3), we cast the problem of estimating Θ into a series of weighted classifications. Specifically we propose the tensor estimate using the sign

representation,

$$\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}\hat{\mathcal{Z}}_\pi, \tag{7}$$

where $\hat{\mathcal{Z}}_\pi \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the π-weighted classifier estimated at levels $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$,

$$\hat{\mathcal{Z}}_\pi = \underset{\mathcal{Z}:\, \mathrm{rank}\mathcal{Z} \leq r}{\arg\min}\; L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi). \tag{8}$$

Here $L(\cdot, \cdot)$ denotes the weighted classification objective defined in (4), where we have plugged $\bar{\mathcal{Y}}_\Omega = (\mathcal{Y}_\Omega - \pi)$ in the expression, and the rank constraint follows from Proposition 3. 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). Step (8) corresponds Figure 2c while (7) to Figure 2d .

The next theorem establishes the statistical convergence for the sign tensor estimate (8), which is an important ingredient for the final signal tensor estimate $\hat{\Theta}$ in (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 (8), $d_{\max} = \max_{k \in [K]} d_k$, and $|\Omega| \geq d_{\max}r$ Then, for all $\pi \in [-1, 1]$ except for a finite number of levels, with very high probability over $\mathcal{Y}_\Omega$,*

$$\mathrm{MAE}(\mathrm{sgn}\hat{\mathcal{Z}}_\pi, \mathrm{sgn}(\Theta - \pi)) \lesssim \left(\frac{d_{\max}r}{|\Omega|}\right)^{\frac{\alpha}{\alpha+2}} + \frac{1}{\rho^2(\pi, \mathcal{N})} \frac{d_{\max}r}{|\Omega|}. \tag{9}$$

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

Combining the sign representability of the signal tensor and the sign estimation accuracy, we obtain the main results on our nonparametric tensor estimation method.

Theorem 3 (Tensor estimation error)**.** *Consider the same conditions of Theorem 2. Let $\hat{\Theta}$ be the estimate in (7). With very high probability over $\mathcal{Y}_\Omega$,*

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim \left(\frac{d_{\max}r}{|\Omega|}\right)^{\frac{\alpha}{\alpha+2}} + \frac{1}{H} + \frac{H d_{\max}r}{|\Omega|}. \tag{10}$$

In particular, setting $H \asymp \left(\frac{|\Omega|}{d_{\max}r}\right)^{1/2}$ yields the error bound

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim \left(\frac{d_{\max}r}{|\Omega|}\right)^{\frac{\alpha}{\alpha+2} \vee \frac{1}{2}}. \tag{11}$$

Theorem 3 demonstrates the convergence rate of our tensor estimation. The bound (10) 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 (11) is generally no better than the corresponding sign error (9). 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 $d_k = d, k \in [K]$, our signal estimate achieves convergence

$$\text{MAE}(\hat{\Theta}, \Theta) \leq \left(\frac{r}{d^{K-1}} \right)^{\frac{\alpha}{\alpha+2} \vee \frac{1}{2}}.$$

Compared to earlier methods, our estimation accuracy applies to both low- and high-rank signal tensors. The rate depends on the sign complexity $\Theta \in \mathscr{P}_{\text{sgn}}(r)$, 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 (TBM)**.** Consider a tensor block model with in total r multiway blocks. Our result implies a rate $\mathcal{O}(d^{-(K-1)/2})$ by taking $\alpha = \infty$. This rate agrees with the previous root-mean-square error (RMSE) for block tensor estimation (Wang and Zeng, 2019).

Example 3 (GLM)**.** 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 marginal density of $\Theta(\omega)$ is bounded as $d \to \infty$. Applying our results with $\alpha = 1$ yields $\mathcal{O}(d^{-(K-1)/3})$. This rate is slightly slower than the parametric RMSE rate (Zhang and Xia, 2018; Wang and Li, 2020). One possible reason is that our estimate remains valid for unknown g and general high-rank tensors with $\alpha = 1$. The nonparametric rate is the price one has to pay for not knowing the form $\Theta = g(\mathcal{Z})$ as a priori.

The following sample complexity for nonparamtric tensor completion is a direct consequence of Theorem 3.

Corollary 1 (Sample complexity for nonparametric completion)**.** *Under the same conditions of Theorem 3 with $\alpha \neq 0$, with high probability over $\mathcal{Y}_\Omega$,*

$$\text{MAE}(\hat{\Theta}, \Theta) \to 0, \quad as \quad \frac{|\Omega|}{d_{\max} r} \to \infty.$$

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 r but not the nonparametric complexity α. Note that $\tilde{\mathcal{O}}(d_{\max} r)$ roughly matches the degree of freedom of sign tensors, suggesting the optimality of our sample requirements.

4.2 Numerical implementation

This section addresses the practical implementation of our estimation (7) illustrated in Figure 2. Our sign representation of the signal estimate $\hat{\Theta}$ is an average of $2H + 1$ sign tensors, which can be solved in a divide-and-conquer fashion. Briefly, we estimate the sign tensors $\mathcal{Z}_\pi$ (detailed in the next paragraph) for the series $\pi \in \mathcal{H}$ through parallel implementation, and then we aggregate the results to yield the output. The estimate enjoys low computational cost similar to a single sign tensor estimation.

For the sign tensor estimation (8), the problem reduces to binary tensor decomposition with a weighted classification loss. A number of algorithms have been developed for this problem (Ghadermarzy et al., 2018; Wang and Li, 2020; Hong et al., 2020). We adopt similar ideas by tailoring

the algorithms to our contexts. Following the common practice in classification, we replace the binary loss $\ell(z, y) = |\text{sgn}z - \text{sgn}y|$ with a surrogate loss $F(m)$ using a continuous function of margin $m := z\text{sgn}(y)$. Examples of large-margin loss are hinge loss $F(m) = (1 - m)_+$, logistic loss $F(m) = \log(1 + e^{-m})$, and nonconvex ψ-loss $F(m) = 2\min(1, (1 - m)_+)$ with $m_+ = \max(m, 0)$. We implement the hinge loss and logistic loss in our algorithm, although our framework is applicable to general large-margin losses (Bartlett et al., 2006).

Algorithm 1 Nonparametric tensor completion

Input: Noisy and incomplete data tensor $\mathcal{Y}_\Omega$, rank r, resolution parameter H.
1: **for** $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$ **do**
2:      Random initialization of tensor factors $\boldsymbol{A}_k = [\boldsymbol{a}_1^{(k)}, \ldots, \boldsymbol{a}_r^{(k)}] \in \mathbb{R}^{d_k \times r}$ for all $k \in [K]$.
3:      **while** not convergence **do**
4:          **for** $k = 1, \ldots, K$ **do**
5:              Update $\boldsymbol{A}_k$ while holding others fixed: $\boldsymbol{A}_k \leftarrow \arg\min_{\boldsymbol{A}_k \in \mathbb{R}^{d_k \times r}} \sum_{\omega \in \Omega} |\mathcal{Y}(\omega) - \pi| F(\mathcal{Z}(\omega)\text{sgn}(\mathcal{Y}(\omega) - \pi))$, where $F(\cdot)$ is the large-margin loss, and $\mathcal{Z} = \sum_{s \in [r]} \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}$ is a rank-r tensor.
6:          **end for**
7:      **end while**
8:      Return $\mathcal{Z}_\pi \leftarrow \sum_{s \in [r]} \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}$.
9: **end for**
Output: Estimated signal tensor $\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \text{sgn}(\mathcal{Z}_\pi)$.

The rank constraints in the optimization (8) have been extensively studied in literature. Recent developments involve convex norm relaxation (Ghadermarzy et al., 2018) and nonconvex optimization (Wang and Li, 2020; Han et al., 2020). Unlike matrices, computing the tensor convex norm is NP hard, so we choose (non-convex) alternating optimization due to its numerical efficiency. Briefly, we use the rank decomposition (2) of $\mathcal{Z} = \mathcal{Z}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K)$ to optimize the unknown factor matrices $\boldsymbol{A}_k = [\boldsymbol{a}_1^{(k)}, \ldots, \boldsymbol{a}_r^{(k)}] \in \mathbb{R}^{d_k \times r}$, where we choose to collect tensor singular values into $\boldsymbol{A}_K$. We numerically solve (7) by optimizing one factor $\boldsymbol{A}_k$ at a time while holding others fixed. Each suboptimization reduces to a convex optimization with a low-dimensional decision variable. Following common practice in tensor optimization (Anandkumar et al., 2014; Hong et al., 2020), we run the optimization from multiple initializations to locate a final estimate with the lowest objective value. The full procedure is described in Algorithm 1.

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.

Figure 3 compares the estimation error under full observation. The MAE decreases with tensor
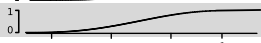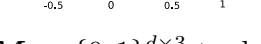
Simulation	Signal Tensor Θ	Rank	Sign Rank	Global α	CDF of Tensor Entries	Noise
1	$\mathcal{C} \times \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \boldsymbol{M}_3$	3^3	$\leq 3^3$	∞		Uniform $[-0.3, 0.3]$
2	$\lvert \boldsymbol{a} \otimes \mathbf{1} \otimes \mathbf{1} - \mathbf{1} \otimes \boldsymbol{a} \otimes \mathbf{1} \rvert$	d	≤ 3	1		Normal $\mathcal{N}(0, 0.15)$
3	$\log(0.5 + Z_{\max})$	$\geq d$	2	1		Uniform $[-0.1, 0.1]$
4	$2.5 - \exp(\mathcal{Z}_{\min}^{1/3})$	$\geq d$	2	1		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.

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.
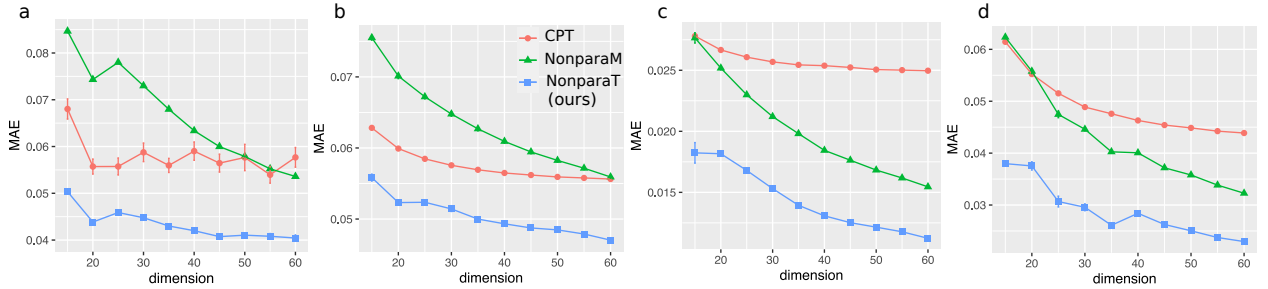

Figure 3: Estimation error versus tensor dimension. Panels (a)-(d) correspond to simulation models 1-4 in Table 1.
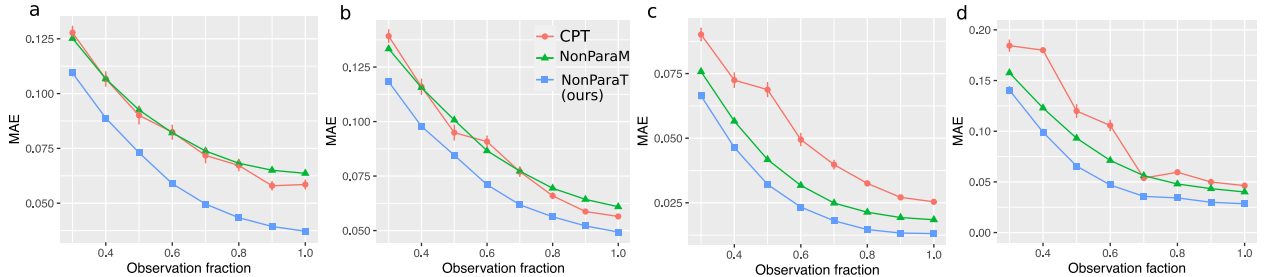

Figure 4: Completion error versus observation fraction. Panels (a)-(d) correspond to simulation models 1-4 in Table 1.

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: 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.

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 $\mathcal{Y} \in \{0, 1\}^{d_1 \times \cdots \times d_K}$) whereas CPT estimation may fall outside the valid range $[0, 1]$.

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. The reported MAEs are averaged over five runs of cross-validation, 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 $H = 20$.

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 $\hat{\Theta} \in \mathbb{R}^{68\times68\times114}$ using 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 and parietal 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. Reported MAEs are averaged over five runs of cross-validation, with 20% entries for testing and 80% for training. 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.
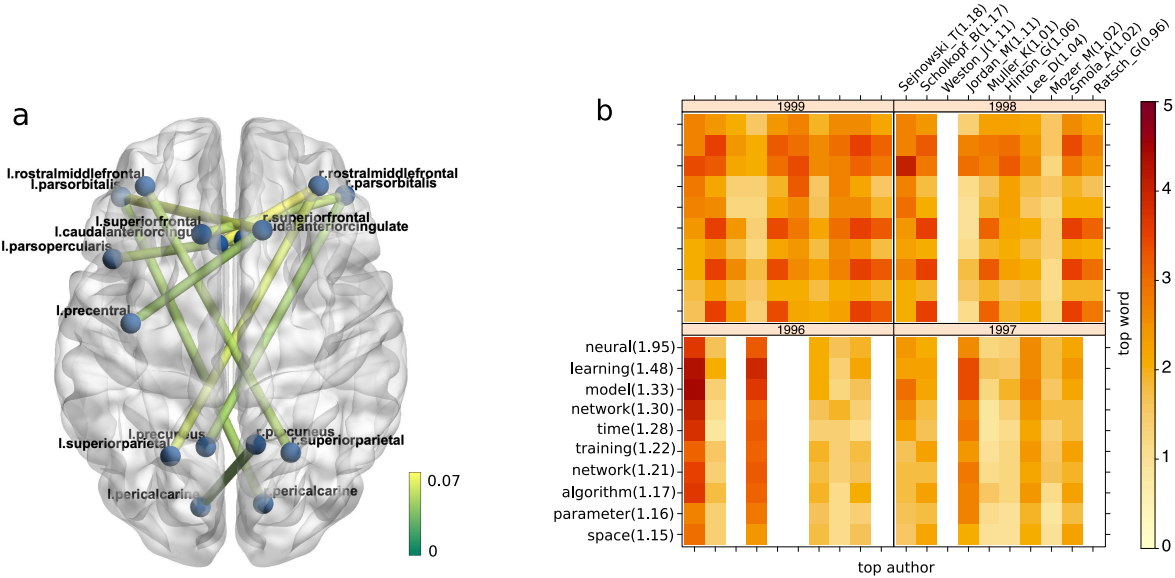

Figure 6: Estimated signal tensors in the data analysis. (a) top edges associated with IQ scores in the brain connectivity data. The color indicates the estimated IQ effect size. (b) top authors and words for years 1996-1999 in the NIPS data. The color indicates the estimated tensor value.

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 provide 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.

7.1 Sensitivity of tensor rank to monotonic transformations

In Section 1, 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.,

$$\hat{r}(\Theta) = \min \left\{ s \in \mathbb{N}_+ \colon \min_{\hat{\Theta} \colon \text{rank}(\hat{\Theta}) \leq s} \frac{\|\Theta - \hat{\Theta}\|_F}{\|\Theta\|_F} \leq 0.1 \right\}.$$

We compute $\hat{r}(\Theta)$ by searching over $s \in \{1, \dots, 30^2\}$, 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 section 3.1, 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 $(d, \dots, d)$-dimensional tensor.

Example 6 (Max hypergraphon)**.** Suppose the tensor Θ takes the form

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

Then

$$\text{rank}(\Theta) \geq d, \text{ and } \text{srank}(\Theta - \pi) \leq 2 \text{ for all } \pi \in \mathbb{R}.$$

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

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

where Θ_i denotes the i-th row of Θ. Now it suffices to show $\mathrm{srank}(\Theta - \pi) \leq 2$ 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 $\log(1 + \frac{i^*-1}{d}) < \pi \leq \log(1 + \frac{i^*}{d})$. 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 $K \geq 3$. By definition of the tensor rank, the rank of a tensor is lower bounded by the rank of its matrix slice. So we have $\mathrm{rank}(\Theta) \geq \mathrm{rank}(\Theta(: , : , 1, \ldots, 1)) = d$. For the sign rank with feasible π, notice that the sign tensor $\mathrm{sgn}(\Theta - \pi)$ takes the similar form as in (12),

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

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 $\mathrm{sgn}(\Theta - \pi) = -2\boldsymbol{a}^{\otimes K} + 1$, where $\boldsymbol{a} = (1, \ldots, 1, 0, \ldots, 0)^T$ takes 1 on the i-th entry if $i < i^*$ and 0 otherwise. Henceforth $\mathrm{srank}(\Theta - \pi) = 2$. $\qquad\square$

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

Proposition 4 (Min/Max hypergraphon). *Let $\mathcal{Z}_{\max} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote a tensor with entries*

$$
\mathcal{Z}_{\max}(i_1, \ldots, i_K) = \max(x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)}), \tag{14}
$$

where $x_{i_k}^{(k)} \in [0, 1]$ are given numbers for all $i_k \in [d_k]$. Let $g \colon \mathbb{R} \to \mathbb{R}$ be a continuous function and $\Theta := g(\mathcal{Z}_{\max})$ be the transformed tensor. For a given $\pi \in [-1, 1]$, suppose the function $g(z) = \pi$ has at most $r \geq 1$ distinct real roots. Then, the sign rank of $(\Theta - \pi)$ satisfies

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

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

Proof. 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

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

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

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

Therefore, $\text{srank}(\Theta - \pi) \leq \text{rank}(\text{sgn}(\Theta - \pi)) = 2$.
- 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 $\mathcal{I} = \{s \in \mathbb{N}_+ : \text{the interval } (z_s, z_{s+1}) \text{ in which } g(z) < \pi\}$. We now prove that the sign tensor $\text{sgn}(\Theta - \pi)$ has rank bounded by $2r - 1$. To see this, consider the tensor indices for which $\text{sgn}(\Theta - \pi) = -1$,

$$
\begin{aligned}
\{\omega : \Theta(\omega) - \pi < 0\} &= \{\omega : g(\mathcal{Z}_{\max}(\omega)) < \pi\} \\
&= \cup_{s \in \mathcal{I}} \{\omega : \mathcal{Z}_{\max}(\omega) \in (z_s, z_{s+1})\} \\
&= \cup_{s \in \mathcal{I}} \Big(\{\omega : x_{i_k}^{(k)} < z_{s+1} \text{ for all } k \in [K]\} \cap \{\omega : x_{i_k}^{(k)} \leq z_s \text{ for all } k \in [K]\}^c \Big).
\end{aligned}
\tag{15}
$$

The equation (15) is equivalent to

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

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

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

where we have denoted the two binary vectors

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

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

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

Combining two cases yields that $\text{srank}(\Theta - \pi) \leq 2r$ for any $r \geq 1$. $\qquad\qquad\square$

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

Proposition 5. *Let $M \in \mathbb{R}^{d_1 \times d_2}$ be a matrix. For any given $K \geq 3$, define an order-K tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ by*

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

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

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

Proof. The conclusion directly follows from the definition of tensor rank. $\qquad\square$

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

$$\mathrm{rank}(M) = d, \quad and \quad \mathrm{srank}(M - \pi) \leq 3, \quad for \ all \ \pi \in \mathbb{R}.$$

Proof. Note that M is a banded matrix with entries

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

Elementary row operation directly shows that M is full rank as follows,

$$\begin{pmatrix} (M_1 + M_d)/(d-1) \\ M_1 - M_2 \\ M_2 - M_3 \\ \vdots \\ M_{d-1} - M_d \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 & 1 \\ -1 & 1 & 1 & \ldots & 1 & 1 \\ -1 & -1 & 1 & \ldots & 1 & 1 \\ \vdots & & & & & \\ -1 & -1 & -1 & \ldots & -1 & 1 \end{pmatrix}.$$

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

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

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

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

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

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

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

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

Example 8 (Stacked identity matrices). Let I be a d-by-d identity matrix. Then

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

Proof. Depending on the value of π, the sign matrix $\text{sgn}(\boldsymbol{I} - \pi)$ falls into one of the three cases: 1) $\text{sgn}(\boldsymbol{I} - \pi)$ is a matrix of all 1; 2) $\text{sgn}(\boldsymbol{I} - \pi)$ is a matrix of all -1; 3) $\text{sgn}(\boldsymbol{I} - \pi) = 2\boldsymbol{I} - \mathbb{1}_d \otimes \mathbb{1}_d$. The former two cases are trivial, so it suffices to show $\text{srank}(\boldsymbol{I} - \pi) \le 3$ in the third case.

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

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

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

7.3 Proofs

7.3.1 Proofs of Propositions 1-3

Proof of Proposition 1. 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 $x \ge 0$, then $g(x) \ge g(0) = 0$. Conversely, if $g(x) \ge 0 = g(0)$, then applying g^{-1} to both sides gives $x \ge 0$. When g is strictly decreasing, the mapping $x \mapsto g(x)$ is sign reversing. Specifically, if $x \ge 0$, then $g(x) \le g(0) = 0$. Conversely, if $g(x) \ge 0 = g(0)$, then applying g^{-1} to both sides gives $x \le 0$. Therefore, $\Theta \simeq g(\Theta)$, or $\Theta \simeq -g(\Theta)$. Since constant multiplication does not change the tensor rank, we have $\text{srank}(\Theta) = \text{srank}(g(\Theta)) \le \text{rank}(g(\Theta))$. $\qquad\square$

Proof of Proposition 2. See Section 7.2 for constructive examples. $\qquad\square$

Proof of Proposition 3. Fix $\pi \in [-1, 1]$. Based on the definition of classification loss $L(\cdot, \cdot)$, the function $\text{Risk}(\cdot)$ 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

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

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

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

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

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

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

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

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

19

7.3.2 Proof of Theorem 1

Proof of Theorem 1. Fix $\pi \in [-1, 1]$. Based on (22) in Proposition 3 we have

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

The Assumption 1 states that

$$\mathbb{P}\left(|\bar{\Theta}| \leq t\right) \leq ct^{\alpha}, \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}). \tag{24}$$

Without future 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 $0 \leq t < \rho(\pi, \mathcal{N})$,

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

  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},

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

  The corresponding lower bound of the inequality (25) becomes

  $$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) \geq \begin{cases} c_1\rho(\pi, \mathcal{N})\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}), & \text{if MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) > 2c(1+\alpha)\rho^{\alpha}(\pi, \mathcal{N}), \\ c_2\left[\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta})\right]^{\frac{1+\alpha}{\alpha}}, & \text{if MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) \leq 2c(1+\alpha)\rho^{\alpha}(\pi, \mathcal{N}), \end{cases}$$

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

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

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

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

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

  $\square$

Remark 2. The proof of Theorem 1 shows that, under Assumption 1,

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

for all $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_R}$. 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).

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

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

$$
= \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}), \tag{28}
$$

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, π)-smoothness of Θ, we have

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

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

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

where the second line comes from the boundedness of classification loss $L(\cdot, \cdot)$, and the third line comes from the inequality $||a - b| - |c - b|| \leq |a - b|$ 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). $\qquad\square$

Proof of Theorem 2. Fix $\pi \in [-1, 1]$. 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 $\|\mathcal{Z}\|_F \leq 1$ without loss of generality. Specifically, let

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

We next apply the empirical process theory to bound $\hat{\mathcal{Z}}$. To facilitate the analysis, we view the data $\mathcal{Y}_\Omega = \{\bar{\mathcal{Y}}(\omega): \omega \in \Omega\}$ 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

$$L(\mathcal{Z},\bar{\mathcal{Y}}_\Omega)-L(\bar\Theta,\bar{\mathcal{Y}}_\Omega)=\frac{1}{n}\sum_{i=1}^n\underbrace{\left[\ell_i(\mathcal{Z},\bar{\mathcal{Y}})-\ell_i(\bar\Theta,\bar{\mathcal{Y}})\right]}_{\overset{\text{def}}{=}\Delta_i(f_{\mathcal{Z}},\bar{\mathcal{Y}})}, \tag{31}$$

is an empirical process induced by function $f_{\mathcal{Z}}\in\mathcal{F}_{\mathcal{T}}$ where $\mathcal{T}=\{\mathcal{Z}\colon\text{rank}(\mathcal{Z})\le r,\ \|\mathcal{Z}\|_F\le1\}$. Note that there is an one-to-one correspondence between sets $\mathcal{F}_{\mathcal{T}}$ and $\mathcal{T}$.

Our remaining proof adopts the techniques of Wang et al. (2008, Theorem 3) to bound (31) over the function family $f_{\mathcal{Z}}\in\mathcal{F}_{\mathcal{T}}$. We summarize only the key difference here but refer to (Wang et al., 2008) for complete proof. Based on Lemma 1, the (α,π)-smoothness of Θ implies

$$\text{Var}\Delta_i(f_{\mathcal{Z}},\bar{\mathcal{Y}})\lesssim\left[\mathbb{E}\Delta_i(f_{\mathcal{Z}},\bar{\mathcal{Y}})\right]^{\frac{\alpha}{1+\alpha}}+\frac{1}{\rho}\mathbb{E}\Delta_i(f_{\mathcal{Z}},\bar{\mathcal{Y}}),\quad\text{for all }f_{\mathcal{Z}}\in\mathcal{F}_{\mathcal{T}}. \tag{32}$$

Applying local iterative techniques in Wang et al. (2008, Theorem 3) to the empirical process (31) with the variance-to-mean relationship (32) gives that

$$\mathbb{P}\left(\text{Risk}(\hat{\mathcal{Z}})-\text{Risk}(\bar\Theta)\ge L_n\right)\lesssim\exp(-nL_n), \tag{33}$$

where the convergence rate $L_n>0$ is determined by the solution to the following inequality,

$$\frac{1}{L_n}\int_{L_n}^{\sqrt{L_n^{\alpha/(\alpha+1)}+\frac{L_n}{\rho}}}\sqrt{\mathcal{H}_{[\,]}(\varepsilon,\mathcal{F}_{\mathcal{T}},\|\cdot\|_2)}d\varepsilon\le C\sqrt{n}, \tag{34}$$

for some constant $C>0$. In particular, the smallest L_n satisfying (34) yields the best upper bound of the error rate. Here $\mathcal{H}_{[\,]}(\varepsilon,\mathcal{F}_{\mathcal{T}},\|\cdot\|_2)$ denotes the L_2-metric, ε-bracketing number (c.f. Definition 3) of family $\mathcal{F}_{\mathcal{T}}$.

It remains to solve for the smallest possible L_n in (34). Based on Lemma 2, the inequality (34) is satisfied with

$$L_n\asymp t_n^{(\alpha+1)/(\alpha+2)}+\frac{1}{\rho}t_n,\quad\text{where }t_n=\frac{d_{\max}rK\log K}{n}.$$

Therefore, by (33), with very high probability.

$$\text{Risk}(\hat{\mathcal{Z}})-\text{Risk}(\bar\Theta)\le t_n^{(\alpha+1)/(\alpha+2)}+\frac{1}{\rho}t_n.$$

Inserting the above bound into (27) gives

$$\text{MAE}(\text{sgn}\hat{\mathcal{Z}},\text{sgn}\bar\Theta)\lesssim[\text{Risk}(\hat{\mathcal{Z}})-\text{Risk}(\bar\Theta)]^{\alpha/(\alpha+1)}+\frac{1}{\rho}[\text{Risk}(\hat{\mathcal{Z}})-\text{Risk}(\bar\Theta)]$$

$$\lesssim t_n^{\alpha/(\alpha+2)}+\frac{1}{\rho^{\alpha/\alpha+1}}t_n^{\alpha/(\alpha+1)}+\frac{1}{\rho}t_n^{(\alpha+1)/(\alpha+2)}+\frac{1}{\rho^2}t_n$$

$$\le 4t_n^{\alpha/(\alpha+2)}+\frac{4}{\rho^2}t_n, \tag{35}$$

where the last line follows from the fact that $a(b^2+b^{(\alpha+2)/(\alpha+1)}+b+1)\le4a(b^2+1)$ with $a=\frac{t_n}{\rho^2}$ and $b=\rho t_n^{-1/(\alpha+2)}$. We plug t_n into (35) and absorb the term $K\log K$ into the constant. The conclusion is then proved. $\qquad\square$

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

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

and

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

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* $\mathcal{T} = \{\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \text{rank}(\mathcal{Z}) \leq r, \|\mathcal{Z}\|_F \leq 1\}$ *and the induced function family* $\mathcal{F}_{\mathcal{T}} = \{f_{\mathcal{Z}} : \mathcal{Z} \in \mathcal{T}\}$. *Set*

$$L_n \asymp \left(\frac{d_{\max} r K \log K}{n}\right)^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi, \mathcal{N})}\left(\frac{d_{\max} r K \log K}{n}\right).$$

Then, the following inequality is satisfied.

$$\frac{1}{L_n} \int_{L_n}^{\sqrt{L_n^{\alpha/(\alpha+1)} + \frac{L_n}{\rho(\pi, \mathcal{N})}}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}_{\mathcal{T}}, \|\cdot\|_2)} d\varepsilon \leq C n^{1/2}, \tag{36}$$

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

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

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

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

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 (36) gives

$$g(L) = \frac{1}{L} \int_L^{\sqrt{L^{\alpha/(\alpha+1)} + \rho^{-1}L}} \sqrt{d_{\max} r K \log\left(\frac{K}{\varepsilon}\right)} d\varepsilon. \tag{37}$$

By the monotonicity of the integrand in (37), we bound $g(L)$ by

$$g(L) \leq \frac{\sqrt{d_{\max} r K}}{L} \int_L^{\sqrt{L^{\alpha/(\alpha+1)} + \rho^{-1}L}} \sqrt{\log\left(\frac{K}{L}\right)} d\varepsilon$$

$$\leq \sqrt{d_{\max} r K (\log K - \log L)} \left(\frac{L^{\alpha/(2\alpha+2)} + \sqrt{\rho^{-1}L}}{L} - 1\right)$$

$$\leq \sqrt{d_{\max} r K \log K} \left(\frac{1}{L^{(\alpha+2)/(2\alpha+2)}} + \frac{1}{\sqrt{\rho L}}\right), \tag{38}$$

23

where the second line follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b > 0$. It remains to verify that $g(L_n) \le Cn^{1/2}$ for L_n specified in (36). Plugging L_n into the last line of (38) gives

$$g(L_n) \le \sqrt{d_{\max} r K \log K} \left(\frac{1}{L_n^{(\alpha+2)/(2\alpha+2)}} + \frac{1}{\sqrt{\rho L_n}} \right)$$

$$\le \sqrt{d_{\max} r K \log K} \left(\left[\left(\frac{d_{\max} r K \log K}{n} \right)^{\frac{\alpha+1}{\alpha+2}} \right]^{-\frac{\alpha+2}{2\alpha+2}} + \left[\rho \left(\frac{d_{\max} r K \log K}{\rho n} \right) \right]^{-\frac{1}{2}} \right)$$

$$\le Cn^{1/2},$$

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

7.3.4 Proof of Theorem 3

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

$$\text{MAE}(\hat{\Theta}, \Theta) = \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi \in \Pi} \text{sgn} \hat{Z}_\pi - \Theta \right|$$

$$\le \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi \in \Pi} \left(\text{sgn} \hat{Z}_\pi - \text{sgn}(\Theta - \pi) \right) \right| + \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi \in \Pi} \text{sgn}(\Theta - \pi) - \Theta \right|$$

$$\le \frac{1}{2H+1} \sum_{\pi \in \Pi} \text{MAE}(\text{sgn} \hat{Z}_\pi, \text{sgn}(\Theta - \pi)) + \frac{1}{H}, \qquad (39)$$

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

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

Write $n = |\Omega|$. Now it suffices to bound the first term in (39). We prove that

$$\frac{1}{2H+1} \sum_{\pi \in \Pi} \text{MAE}(\text{sgn} \hat{Z}_\pi, \text{sgn}(\Theta - \pi)) \lesssim t_n^{\alpha/(\alpha+2)} + \frac{1}{H} + H t_n, \quad \text{with } t_n = \frac{d_{\max} r K \log K}{n}. \qquad (40)$$

Theorem 2 implies that the sign estimation accuracy depends on the closeness of $\pi \in \mathcal{H}$ to the mass points in $\mathcal{H}$. Therefore, we partition the level set $\pi \in \mathcal{H}$ based on their closeness to $\mathcal{H}$. Specifically, let $\mathcal{N}_H \overset{\text{def}}{=} \bigcup_{\pi' \in \mathcal{N}} \left(\pi' - \frac{1}{H}, \pi' + \frac{1}{H} \right)$ denote the set of levels at least $\frac{1}{H}$-close to the mass points. We expand (40) by

$$\frac{1}{2H+1} \sum_{\pi \in \Pi} \text{MAE}(\text{sgn} \hat{Z}_\pi, \text{sgn}(\Theta - \pi))$$

$$= \frac{1}{2H+1} \sum_{\pi \in \Pi \cap \mathcal{N}_H} \text{MAE}(\text{sgn} \hat{Z}_\pi, \text{sgn}(\Theta - \pi)) + \frac{1}{2H+1} \sum_{\pi \in \Pi \cap \mathcal{N}_H^c} \text{MAE}(\text{sgn} \hat{Z}_\pi, \text{sgn}(\Theta - \pi)). \qquad (41)$$

By assumption, the first term involves only finite number of summands and thus can be bounded by $4C/(2H+1)$ where $C > 0$ is a constant such that $|\mathcal{N}| \le C$. We bound the second term using the explicit forms of $\rho(\pi, \mathcal{N})$ in the sequence $\pi \in \Pi \cap \mathcal{N}_H^c$. Based on Theorem 2,

$$\frac{1}{2H+1} \sum_{\pi \in \Pi \cap \mathcal{N}_H^c} \text{MAE}(\text{sgn} \hat{Z}_\pi, \text{sgn}(\Theta - \pi)) \lesssim \frac{1}{2H+1} \sum_{\pi \in \Pi \cap \mathcal{N}_H^c} t_n^{\alpha/(\alpha+2)} + \frac{t_n}{2H+1} \sum_{\pi \in \Pi \cap \mathcal{N}_H^c} \frac{1}{\rho^2(\pi, \mathcal{N})}$$

$$\leq t_n^{\alpha/(\alpha+2)} + \frac{t_n}{2H+1} \sum_{\pi \in \Pi \cap \mathcal{N}_H^c} \sum_{\pi' \in \mathcal{N}} \frac{1}{|\pi - \pi'|^2}$$

$$\leq t_n^{\alpha/(\alpha+2)} + \frac{t_n}{2H+1} \sum_{\pi' \in \mathcal{N}} \sum_{\pi \in \Pi \cap \mathcal{N}_H^c} \frac{1}{|\pi - \pi'|^2}$$

$$\leq t_n^{\alpha/(\alpha+2)} + 2CHt_n,$$

where the last inequality follows from the Lemma 3. Combining the bounds for the two terms in (41) completes the proof for conclusion (40). Finally, plugging (40) into (39) yields

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim \left(\frac{d_{\max} r K \log K}{|\Omega|} \right)^{\alpha/(\alpha+2)} + \frac{1}{H} + H \frac{d_{\max} r K \log K}{|\Omega|}.$$

The conclusion follows by absorbing $K \log K$ into the constant term in the statement. $\qquad\square$

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,*

$$\sum_{\pi \in \Pi \cap \mathcal{N}_H^c} \frac{1}{|\pi - \pi'|^2} \leq 4H^2.$$

Proof of Lemma 3. Notice that all points $\pi \in \Pi \cap \mathcal{N}_H^c$ satisfy $|\pi - \pi'| > \frac{1}{H}$ for all $\pi' \in \mathcal{N}$. We use this fact to compute the sum

$$\sum_{\pi \in \Pi \cap \mathcal{N}_H^c} \frac{1}{|\pi - \pi'|^2} = \sum_{\frac{h}{H} \in \Pi \cap \mathcal{N}_H^c} \frac{1}{|\frac{h}{H} - \pi'|^2}$$

$$\leq 2H^2 \sum_{h=1}^H \frac{1}{h^2}$$

$$\leq 2H^2 \left\{ 1 + \int_1^2 \frac{1}{x^2} dx + \int_2^3 \frac{1}{x^2} dx + \cdots + \int_{H-1}^H \frac{1}{x^2} dx \right\}$$

$$= 2H^2 \left(1 + \int_1^H \frac{1}{x^2} dx \right) \leq 4H^2,$$

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

8 Conclusion

We have developed a tensor completion method that addresses both low- and high-rankness based on sign series representation. Our work provide a nonparametric framework for tensor estimation, and we obtain results previously impossible. We hope the work opens up new inquiry that allows more researchers to contribute to this field.

Acknowledgements

References

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

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.

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.

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. 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.

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

Wang, J., Shen, X., and Liu, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167.

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.

Xu, J. (2018). Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, pages 5433–5442.

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.