Journal of Machine Learning Research X (20XX) 1-51 Submitted 6/2022; Published XX/XX

Beyond low-rankness: Nonparametric estimation and completion for sign-representable tensors

Chanwoo Lee CHANWOO.LEE@WISC.EDU
Department of Statistics
University of Wisconsin-Madison

Miaoyan Wang MIAOYAN.WANG@WISC.EDU
Department of Statistics
University of Wisconsin-Madison

Editor:

Abstract

Rapid developments in modern technologies have made tensor data readily available in daily life. Existing tensor estimation methods often rely on three assumptions: (i) a global low-rank structure across all tensor entries; (ii) a known, linear relationship between observed space and latent low-rank representation; (iii) separate treatments for tensors under various noise distributions. Here, we propose a nonparametric tensor estimation method based on a new model coined as *sign representable tensors*. Our tensor model efficiently addresses possibly high-rank signals, allows various data types, and enjoys invariant property under unknown order-preserving entrywise transformations. We establish the excess risk bound, estimation error rate, and sample complexity for the tensor estimation problem with missingness. A key theoretical contribution is a provable reduction of complex (continuous) parameter estimation to a series of simpler (binary) sign estimations; this learning-reduction paradigm can be of independent interest. Further, we develop a divide-and-conquer algorithm to nonparametric tensor estimation with accuracy guarantees. 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.

Keywords: tensor estimation, tensor completion, nonparametric statistics, weighted classification, rate of convergence, sample complexity.

1. Introduction

1.1 Motivation

Higher-order tensors have recently received much attention in enormous fields including social networks (Anandkumar et al., 2014), neuroscience (Zhang et al., 2019), and genomics (Hore et al., 2016). Tensor methods provide effective representation of the hidden structure in multiway data. Our starting point is the signal plus noise model,

$$\mathcal{Y} = \Theta + \mathcal{E}, \quad \text{with} \quad \mathbb{E}(\mathcal{E}) = 0, \tag{1}$$

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 zero-mean noise tensor. Here, we take the convention that $\mathbb{E}(\cdot)$ is applied to the tensor $\mathcal{E}$ in an entrywise fashion. The goal is to estimate Θ from the noisy observation $\mathcal{Y}$ with possibly missing entries. In particular, we focus on the following two problems:

Q1 (Nonparametric tensor estimation). How to accurately 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 *under a wide range of observation models*?

There is a huge literature on structured tensor estimation based on low-rankness of the signal tensor Θ (Jain and Oh, 2014; Montanari and Sun, 2018; Anandkumar et al., 2014; Allen, 2012) . 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). Generalized linear tensor models (Wang and Li, 2020; Hu et al., 2022) are also proposed by assuming the low-rankness up to *known* entrywise transformation. 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 low-rank models.

Sensitivity of tensor rank to transformation. 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 $\text{rank}(\mathcal{Z}) = 3$ (formal rank 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. For illustration, we plot the numerical CP and Tucker rank of Θ versus c. The experiment details are provided in Appendix A. Figure 1a shows that $\text{rank}(\Theta)$ far exceeds the intrinsic rank $\text{rank}(\mathcal{Z}) = 3$; in fact, the rank increases rapidly with the transformation level c. This observation suggests the failure of $\text{rank}(\Theta)$ in capturing the intrinsic model complexity. In practical applications of digital processing (Ghadermarzy et al., 2018) and genomics analysis (Hore et al., 2016), the tensor of interest often undergoes *unknown* entrywise transformation prior to measurements. The rank sensitivity renders traditional low-rank tensor methods less desirable in the presence of even mild order-preserving nonlinearities.

Limitation of low-rank family. 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), and here we extend the structure to higher-order tensors. We find that neither Θ nor $\mathcal{Z}$ is low-rank; in fact, Figure 1b shows that all top d tensor singular values are well above zero. Unlike matrices, tensor rank may even exceed the dimension d. This example shows the failure of classical low-rank models in capturing this type of tensor structures.

In this paper, we develop a *nonparametric tensor model* that efficiently addresses a wide range of tensor estimation problems. In the above and many other examples, the signal tensor Θ is possibly high rank. Classical low-rank models will miss important structures. The observations have motivated us to develop a more flexible tensor estimation method.

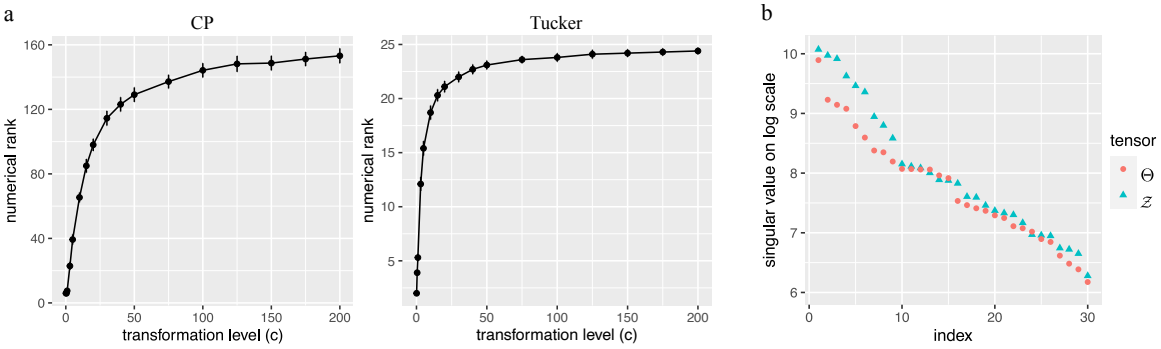


Figure 1: (a) Numerical tensor rank (CP and Tucker) vs. transformation level c in the first example. For Tucker rank, we plot the rank along first mode for illustration. (b) Top $d = 30$ tensor singular values in the second example. The details of experiment are provided in Appendix A.

We will revisit these two motivating examples in Section 2.3, and show that our method overcomes the aforementioned limitations.

1.2 Our contributions

We develop a nonparametric tensor model to address the aforementioned challenges. Existing tensor estimation methods often rely on three assumptions: (i) a global low-rank structure across all tensor entries; (ii) a known, linear relationship between observed space and latent low-rank representation; (iii) separate treatments for tensors under various noise distributions. By contrast, our proposed nonparametric model (i) enjoys rank invariance under monotonic transformations, (ii) allows both low-rank and high-rank effects, and (iii) provides a unified framework for both continuous and discrete data types.

We highlight four main contributions that set our work apart from earlier literature.

In modeling perspective, we present an innovative *nonparametric* notion of complexity that incorporates more flexible structure than previously possible. Existing estimation theory (Anandkumar et al., 2014; Montanari and Sun, 2018; Cai et al., 2019) mostly focuses on the regime of low rank 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. Our key crux 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 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.

In statistical perspective, we develop a *provable learning-reduction paradigm* by connecting the tensor estimation problem to a series of weighted classification tasks. This characterization converts a difficult $\mathbb{R}$-valued estimation problem, "*what* is the value of Θ at index ω?", to a series of relatively simpler $\{0,1\}$-valued problems, "*whether or not* the value of Θ at index ω falls below a threshold?" We establish the excess risk bounds, estimation error rates, and sample complexity. As a by-product, the learning-reduction approach relaxes the common assumptions of homoscedastic normal noise; this flexibility provides a unified treatment for Gaussian, Bernoulli, and Binomial tensors. The result can be of independent interest for general nonparametric problems. Considering the incisive minds that have studied tensor completion and nonparametric problems separately, our paper provides such connections that were not previously formulated.

In computational perspective, a number of efficient algorithms are readily available for $\{0,1\}$-valued tensor problems (Wang and Li, 2020; Han et al., 2022; 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. We develop a divide-and-concur approach to efficiently combine existing algorithms, thereby achieving computational accuracy without the need to reinvent the wheel. The flexibility to import and adapt existing algorithms is one advantage of our method.

In application perspective, we demonstrate the advantages of our method through simulations and data applications. We apply our method to two datasets, one on brain connectivity study and the other on topic data analysis. The result shows that our approach not only improves the accuracy over previous tensor methods but also shows the robustness to model misspecification. Our method will help the practitioners efficiently analyze tensor datasets in various areas. Toward this end, our have released the software package at CRAN[1].

1.3 Related work

Our work is closely related to but also clearly distinctive from several lines of existing research. We review related literature for comparison.

High-rank matrix models. High-rank matrix estimation has been extensively studied under graphon models (Chatterjee, 2015; Zhang et al., 2017), nonlinear models (Ganti et al., 2015) and subspace clustering (Ongie et al., 2017; Fan and Udell, 2019). In particular, the recent work (Lee et al., 2021) proposes a general nonparametric framework to address a variety of matrix problems including regression and completion. However, high-rank tensor problems are more challenging for at least two reasons: (i) the tensor rank often exceeds the dimension when order K greater than two (Anandkumar et al., 2017); this is in sharp contrast to matrices ($K = 2$); (ii) the lack of tensor analogy for Eckart-Young theorem (Kolda, 2003) poses theoretical challenges to extend matrix SVD thresholding (Chatterjee, 2015; Xu, 2018)

1. https://CRAN.R-project.org/package=TensorComplete

to higher-order tensors. A full exploitation of the higher-order structure is therefore needed; we address this challenge in the paper.

Nonparametric tensor regression. The goal of tensor regression problem is to relate high-dimensional tensor covariates to a scalar response. There have been recent developments of nonparametric methods for tensor regression problems. Hao et al. (2021) proposed the additive regression model that exploits the sparse and low-rank structure in the tensor by extending the usual spline basis function. Zhou et al. (2020) combined multiple broadcasting operation and low-rank scaling coefficients to introduce nonlinearity on tensor entries. However, their results rely on low-rankness of the tensor, which might be restrictive in applications. By contrast, our method applies to both low-rank and high-rank tensors. Additionally, one fundamental difference is our tensor estimation problem works with a single tensor observation while tensor regression problem works with multiple tensor samples and corresponding responses. As a result, our theoretical error bound is based on tensor dimension while that of tensor regression is a function of the sample size.

Graphon and hypergraphon. Graphon is a measurable function representing the limit of a sequence of exchangeable random graphs/matrices (Klopp et al., 2017; Gao et al., 2015; Chan and Airoldi, 2014). Similarly, hypergraphon (Zhao, 2015; Lovász, 2012) is introduced as a limiting function based on a sequence of K-uniform hypergraphs, in which edges can join K vertices with $K \geq 3$. A special case of our *sign representable tensor model* is related to, but also distinctive from, general hypergraphons; see Section 2.3 for details. Both our model and hypergraphon use functions to represent signal tensors. However, there are two remarkable differences. First, unlike the matrix case where graphon is represented as a bivariate function, general hypergraphons for order-K tensors are known to be represented by $(2^K - 1)$-variate function (Zhao, 2015). Our sign series representation depends on K coordinates only, and in this sense, our model shares common ground with *simple hypergraphons* (Balasubramanian, 2021). Second, our functional representation uses deterministic design points, whereas simple hypergraphons use random design points. These two choices lead to a notable difference in the root-mean-square rate $\tilde{\mathcal{O}}(d^{-(K-1)/3})$ (ours) versus. $\mathcal{O}(d^{-1})$ (simple hypergraphon; Balasubramanian (2021)) where d is tensor dimension. We see that our model has a substantially faster rate than simple hypergraphon as tensor order $K \to \infty$. Further comparison will be provided in Sections 4.2 and 5.3.

1.4 Notation and organization

We use $\mathbb{R}$ for the set of real numbers, $\mathbb{N} = \{0, 1, 2, \dots, \}$ for integers, and $\mathbb{N}_+ = \{1, 2, \dots\}$ for non-negative integers. We use $\mathrm{sgn}(\cdot)\colon \mathbb{R} \to \{-1, 1\}$ to denote the sign function, where $\mathrm{sgn}(y) = 1$ if $y \geq 0$ and -1 otherwise. We allow univariate functions (such as $\mathrm{sgn}(\cdot)$, expectation $\mathbb{E}(\cdot)$, function $f\colon \mathbb{R} \to \mathbb{R}$, etc) to be applied to tensors in an element-wise manner. We use $[n] = \{1, \dots, n\}$ for n-set with $n \in \mathbb{N}_+$. We denote $a_n \lesssim b_n$ if $\lim_{n\to\infty} a_n/b_n \leq c$ for some constant $c > 0$, and $a_n \asymp b_n$ if $c_1 \leq \lim_{n\to\infty} a_n/b_n \leq c_2$ for some constants $c_1, c_2 > 0$. We use $\mathcal{O}(\cdot)$ to denote the big-O notation, $\tilde{\mathcal{O}}(\cdot)$ to denote the variant hiding logarithmic factors, and $\mathrm{poly}(d)$ to denote the polynomial complexity in integer d.

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 $\omega \in [d_1] \times \cdots \times [d_K]$. Denote $d_{\max} = \max_{k \in [K]} d_k$. The CP decomposition (Hitchcock, 1927) is defined by

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

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 CP rank, denoted $\text{rank}(\Theta)$. Occasionally, we also use tensor Tucker rank, defined by

$$\text{Tucker-rank}(\Theta) = (r_1, \ldots, r_k), \quad \text{where} \quad r_k = \text{rank}(\text{Unfold}_k(\Theta)) \quad \text{for all } k \in [K],$$

where $\text{Unfold}_k(\Theta)$ denotes the unfolding operation that reshapes the tensor along model k into a matrix of size d_k-by-$\prod_{i \neq k} d_i$, and $\text{rank}(\text{Unfold}_k(\Theta))$ denotes the usual matrix rank. The tensor maximum norm is denoted $\|\Theta\|_\infty = \max_\omega |\Theta(\omega)|$, and the the tensor Frobenius norm is denoted $\|\Theta\|_F = \sqrt{\sum_\omega \Theta^2(\omega)}$. We use $\mathbf{1}$ to denote a vector with all entries 1; the dimension of the vector should be clear from the contexts.

The rest of the paper is organized as follows. In Section 2, we propose the sign representable signal tensor model and a range of noise models. Algebraic and statistical properties of sign representable tensor model are provided in Sections 3 and Section 4, respectively. Section 5 presents the learning-reduction approach to estimation and the finite sample accuracy. Important applications and extensions are also provided. Simulation and two data applications are presented in Section 7. Section 8 concludes the paper with discussion. Proofs and details for experiments are deferred to Appendix.

2. Sign representable tensor model

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

$$\mathcal{Y} = \Theta + \mathcal{E}, \quad \mathbb{E}(\mathcal{E}) = 0, \tag{3}$$

where $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is an unknown signal tensor of interest, and $\mathcal{E} \in \mathbb{R}^{d_1 \times \cdots \times d_k}$ is a noise tensor consisting of zero-mean, independent *but not necessarily identically* distributed entries. The goal is to estimate Θ from $\mathcal{Y}$ with possibly missing entries.

In the following sections, we propose the *sign representable model* for the signal tensor Θ and a range of noise models for $\mathcal{E}$ in (3). We will show that our model addresses both low- and high-rank signals, while encompassing important tensor models—including CP models, Tucker models, single index models, simple hypergraphons—as special cases. In addition, our noise model allows general mean-to-variance dependence, thereby providing a unified framework for analyzing Gaussian, Bernoulli, and Binomial tensors.

2.1 Signal tensor model

We first develop the structure assumptions on Θ in (3). Let Θ be the signal 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.,

$$\mathrm{srank}(\Theta) := \min\{\mathrm{rank}(\Theta') \colon \Theta' \simeq \Theta,\ \Theta' \in \mathbb{R}^{d_1 \times \cdots \times d_K}\}. \tag{4}$$

The concept of sign-rank is firstly introduced in combinatorics (Cohn and Umans, 2013), complexity theory (Alon et al., 2016), and quantum mechanics (De Wolf, 2003). Here we exploit this notion for nonparametric tensor estimation. The sign equivalence $\simeq$ induces a quotient space in $\mathbb{R}^{d_1 \times \cdots \times d_K}$. Note that the sign-rank concerns only the sign pattern but discards the magnitude information of Θ. In particular, $\mathrm{srank}(\Theta) = \mathrm{srank}(\mathrm{sgn}\Theta)$.

We now propose that the signal tensor Θ in model (3) falls into a general tensor family, which we coin as "sign representable tensors".

Definition 2 (Sign representable tensor model)**.** We define the sign-r representable tensor family by

$$\mathscr{P}_{\mathrm{sgn}}(r) = \{\Theta \colon \max_{\pi \in [-1,1]} \mathrm{srank}(\Theta - \pi) \le r,\ \|\Theta\|_\infty \le 1\}. \tag{5}$$

A tensor Θ is called *sign-r representable* if $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$. We call the collection of binary tensors $\{\mathrm{sgn}(\Theta - \pi) \colon \pi \in [-1,1]\}$ the *sign tensor series* corresponding to Θ. In general, the norm bound $\|\Theta\|_\infty \le 1$ in (5) can be replaced by $\|\Theta\|_\infty \le L$, where $L \in \mathbb{R}_+$ is an arbitrary constant; we choose $L = 1$ for notational simplicity.

Remark 1 (Comparison with classical low-rank family)**.** Our sign representable tensor family improves the classical low-rank family. The notion of sign-r representability concerns the *local* structure of Θ. If we interpret the tensor Θ as a function $\Theta : [d_1] \times \cdots \times [d_K] \mapsto [-1,1]$, then the constraint $\mathrm{srank}(\Theta - \pi) \le r$ imposes local model structure at each π in the function range $[-1,1]$. The shifted tensor $(\Theta - \pi)$ needs not to be low-rank itself; only its sign pattern matters. This structure relaxes the classical low-rank family, where a *global* low-rankness is imposed across all entry values of Θ. In Section 2.3, we will show that our proposed sign representable tensor family substantially enriches the classical low-rank family. In particular, both tensor examples in Section 1.1 (see Figure 1) are incorporated in $\mathscr{P}_{\mathrm{sgn}}(r)$ but not in low-rank family.

2.2 Noise tensor model

We now describe our model for noise tensor $\mathcal{E}$ in (3). We require that the noise tensor entries follow independent zero-mean assumption $\mathbb{E}(\mathcal{E}) = 0$ but otherwise no other distribution assumptions. In particular, we allow heterogeneous noise such that the noise variance $\mathrm{Var}(\mathcal{E}(\omega))$ may depend on index ω. Our model (3) therefore provides a unified framework for a wide range of data types, including Gaussian, Bernoulli, Binomial, and mixtures of

them. For instance, our model (3) incorporates the probability tensor estimation problem where $\mathcal{Y}$ is a binary tensor with entries $\{0, 1\}$ from Bernoulli distribution. In such a case, $\Theta = \mathbb{E}(\mathcal{Y})$ represents the probability tensor to be estimated, and the noise variance depends on the mean, $\mathrm{Var}(\mathcal{E}(\omega)) = \Theta(\omega)(1 - \Theta(\omega))$.

We also allow missing data. Our observation is a possibly incomplete data tensor, denoted $\mathcal{Y}_\Omega$, from (3), where $\Omega \subset [d_1] \times \cdots \times [d_K]$ denotes the index set of observed entries. We consider a general model of 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 $\sum_{\omega \in [d_1] \times \cdots \times [d_K]} p_\omega = 1$. Let $\omega \sim \Pi$ denote the sampling rule, meaning that ω's in Ω are i.i.d. draws with replacement from distribution Π. The goal is to estimate Θ from Y_Ω.

2.3 Important examples

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

Example 1 (CP/Tucker low-rank models)**.** The CP and Tucker low-rank tensors are 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., $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r + 1)$ (the constant 1 is due to $\mathrm{rank}(\Theta - \pi) \le r + 1$). Similar results hold for Tucker low-rank tensors $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r + 1)$, 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 $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$.

Example 3 (Generalized linear models (GLMs))**.** Let $\mathcal{Y}$ be a binary tensor from a logistic model (Wang and Li, 2020) with mean $\Theta = \mathrm{logit}(\mathcal{Z})$, 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., $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r + 1)$.

Example 5 (Structured tensors with repeating entries)**.** Here we revisit the model introduced in Figure 1c of Section 1.1. Let Θ be an order-K tensor with entries $\Theta(i_1, \ldots, i_K) = \log(1 + \max_k x^{(k)}_{i_k})$, where $x^{(k)}_{i_k}$ are given numbers in $[0, 1]$ for all $i_k \in [d_k], k \in [K]$. We conclude that $\Theta \in \mathscr{P}_{\mathrm{sgn}}(2)$, because the sign tensor $\mathrm{sgn}(\Theta - \pi)$ with an arbitrary $\pi \in (0,\ \log 2)$ is a block tensor with at most two blocks.

Example 5 extends to general tensors with entries $\Theta(i_1, \ldots, i_K) = g(\max_k x^{(k)}_{i_k})$, where $g(\cdot)$ is an arbitrary 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 Proposition 2 in the next section. Same conclusion holds if the maximum in $g(\cdot)$ is replaced by the minimum.

3. Algebraic properties of sign representable tensors

In this section, we present algebraic properties for sign representable tensors (5). 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. We find that the sign-rank is upper bounded by tensor rank. More generally, we show the following properties.

Proposition 1 (Relationship between sign-rank and usual rank)**.**
(a) [Upper bounds] $\max_{\pi\in[-1,1]} \text{srank}(\Theta - \pi) \le \text{rank}(\Theta) + 1$.
(b) [Invariance under monotonic transformation] If $\Theta \in \mathscr{P}_{\text{sgn}}(r)$, then $g(\Theta)/\left\|g(\Theta)\right\|_\infty \in \mathscr{P}_{\text{sgn}}(r+1)$ for any strictly monotonic function $g\colon \mathbb{R} \to \mathbb{R}$.
(c) [Broadness] For every order $K \ge 2$ and every dimension d, there exist tensors $\Theta \in \mathbb{R}^{d\times\cdots\times d}$ such that $\text{rank}(\Theta) \ge d$ but $\text{srank}(\Theta - \pi) \le 2$ for all $\pi \in \mathbb{R}$.

Proposition 1 highlights the advantages of using sign-rank in the high-dimensional tensor analysis. The property (a) implies that the classical low-rank family is a special case of our sign representable tensor family. The property (b) shows that, compared to classical tensor rank, the sign-rank remains nearly invariant under monotonic transformations; this is because $\text{srank}(g(\Theta)) \le 1 + \text{srank}(\Theta)$ for all strictly monotonic function g. The property (c) shows that the sign-rank can be dramatically smaller than the classical tensor rank. Therefore, our model family $\mathscr{P}_{\text{sgn}}(r)$ is strictly richer than the usual low-rank family.

In Section 2.3, we have provided several tensor examples with high tensor rank but low sign-rank. Here we provide additional examples in which the tensor rank grows with dimension but the sign-rank remains a constant. Important examples are generalized into related propositions. Proofs are deferred to Appendix B. For notational simplicity, we consider tensors of equal dimension at each mode; i.e., $d_1 = d_2 = \cdots = d_K = d$.

Example 6 (Structured tensors with repeating entries)**.** Suppose a 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

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

More generally, Examples 5-6 are special cases of the following proposition.

Proposition 2 (Rank relations for 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 $r \ge 1$ distinct real roots. For given numbers $x_{i_k}^{(k)} \in [0,1]$ with $i_k \in [d_k]$, define a tensor $\Theta \in \mathbb{R}^{d_1\times\cdots\times d_K}$ with entries

$$\Theta(i_1, \ldots, i_K) = g(\max(x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)})), \quad (i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K]. \tag{6}$$

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

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

The same conclusion holds if we use min in place of max in (6).

We conclude this section by providing additional examples in which $\mathrm{rank}(\Theta) \ge d$ but $\mathrm{srank}(\Theta) \le c$ for a constant c independent of d. These examples highlight the advantages of using sign-rank in high-dimensional tensor analysis. We first state an example in the matrix case, i.e, $K = 2$.

Example 7 (Stacked banded matrices)**.** Let $\boldsymbol{a} = (1, 2, \ldots, d)^T$ 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

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

Proposition 3 (Rank relations for stacked banded 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

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

By Example 7 and Proposition 3, we conclude that the stacked banded tensor of the form $\Theta = |\boldsymbol{a} \otimes \mathbf{1}_{d_2} \otimes \cdots \otimes \mathbf{1}_{d_K} - \mathbf{1}_{d_1} \otimes \boldsymbol{a} \otimes \mathbf{1}_{d_3} \otimes \cdots \otimes \mathbf{1}_{d_K}|$ has high usual rank but low sign-rank for general $K \ge 3$. This stacked banded tensor will be used in Simulation Section 7.

4. Statistical properties of sign representable tensors

In this and next sections, we present our main statistical results for sign representable tensors model (3). This section focuses on *population* characterization of sign tensor series $\{\mathrm{sgn}(\Theta - \pi) \colon \pi \in [-1, 1]\}$, including risk function, mean absolute error bound, and uniqueness guarantees. In Section 5, we will provide the *finite sample accuracy* for estimating target signal $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$ from sign tensor series. Table 1 briefly summarizes the main notations for reader's convenience; notations will be detailed when first introduced in the main texts.

We show that the sign tensor series are *uniquely* characterized by *weighted* classification risks under smoothness regularities (Sections 4.1-4.2). The choice of weights turns out to be crucial to address heteroscedasticity in noises (Section 4.3). The results can be of independent interest for general nonparametric problems.

4.1 Classification risk with l_1 weights

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 $\omega \in \Omega$. Our goal is to estimate sign tensor series $\{\mathrm{sgn}(\Theta - \pi) : \pi \in [-1, 1]\}$ associated with

Notation	Definition
$\mathscr{P}_{\text{sgn}}(r)$	r-sign representable tensor family
π	a scalar value in the range $[-1, 1]$
$\bar{\mathcal{Y}}$	π-shifted tensor with entries $\bar{\mathcal{Y}}(\omega) = \mathcal{Y}(\omega) - \pi$
$L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega)$	weighted classification (sample) loss
$\text{Risk}(\mathcal{Z})$	weighted classification (population) risk
Δs	a small tolerance $\Delta s = 1/\prod_{k=1}^{K} d_k$
$\mathcal{N}$	irregular π's where the pseudo density is unbounded
$\lvert\mathcal{N}\rvert_{\text{jump}}$	the covering number of $\mathcal{N}$ with $2\Delta s$-bin's
$\rho(\pi, \mathcal{N})$	an adjusted distance from π to the nearest point in $\mathcal{N}$
$\text{MAE}(\Theta_1, \Theta_2)$	the mean absolute error between two tensors

Table 1: Main notation used in Sections 4-5.

the target signal $\Theta \in \mathscr{P}_{\text{sgn}}(r)$. We propose a *weighted classification loss* function

$$L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \underbrace{|\bar{\mathcal{Y}}(\omega)|}_{l_1 \text{ weight}} \times \underbrace{|\text{sgn}\mathcal{Z}(\omega) - \text{sgn}\bar{\mathcal{Y}}(\omega)|}_{\text{classification loss}}, \tag{7}$$

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 entrywise l_1 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.

Our proposed weighted classification loss (7) is important for characterizing $\text{sgn}(\Theta - \pi)$. Define the *weighted classification risk*

$$\text{Risk}(\mathcal{Z}) = \mathbb{E}_{\pi \sim \Pi} \mathbb{E}_{\mathcal{Y}(\omega)} L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega), \tag{8}$$

where the expectation is taken with respect to $\mathcal{Y}(\omega)$ under model (3) with sampling distribution $\omega \sim \Pi$. The form of $\text{Risk}(\cdot)$ implicitly depends on π; for notational simplicity we suppress π when no confusion arises. The following Theorem 1 ensures that the target signal $(\Theta - \pi)$ falls into the optimizers of risk (8).

Theorem 1 (Weighted classification risk minimizer)**.** Suppose the data $\mathcal{Y}_\Omega$ is generated from model (3) with $\Theta \in \mathscr{P}_{\text{sgn}}(r)$. Then, $(\Theta - \pi)$ minimizes the weighted classification risk (8); that is,

$$\text{Risk}(\Theta - \pi) = \inf\{\text{Risk}(\mathcal{Z}) \colon \mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}\} = \inf\{\text{Risk}(\mathcal{Z}) \colon \text{rank}(\mathcal{Z}) \le r\}. \tag{9}$$

Remark 2 (Noise effects)**.** The proof of Theorem 1 is provided in Appendix C. We summarize the main novelty here. The expectation in the risk is evaluated with respect to the distribution of residual $\mathcal{E} := \mathcal{Y} - \mathbb{E}(\mathcal{Y})$ and the sampling distribution $\omega \sim \Pi$. We find that, when l_1 weights are used, only two noise assumptions are invoked in the risk: (i) entrywise independence of $\mathcal{E}$ (conditional on Θ), and (ii) zero-mean assumption $\mathbb{E}(\mathcal{E}) = 0$. Other aspects of the noise (e.g. variance, shape, distribution) are irrelevant. Therefore, our conclusion (9) holds under a very broad range of noise models, including Gaussian, Bernoulli, Binomial, Poisson residuals. More discussions about the choice of l_1 weights are provided in Section 4.3.

Remark 3 (Landscape of risk minimizers)**.** Theorem 1 ensures that the target signal $(\Theta - \pi)$ is the risk minimizer. However, the converse is false; that is, the risk minimizers may not equal to $(\Theta - \pi)$ due to two types of ambiguities. The first type is the sign equivalence. Based on the definition of weighted classification loss, the risk relies only on the sign pattern of the tensor. The risk minimizers should be interpreted in the quotient space under sign equivalence $\simeq$. That is, all tensors that are sign equivalent to $(\Theta - \pi)$ belong to the risk minimizers. This ambiguity is easy to address, since our goal here is the intermediate sign tensor series $\mathrm{sgn}(\Theta - \pi)$ not $(\Theta - \pi)$. The second type is the possibility of *multiple* solutions to the risk minimization in the quotient space. We shall establish the essential *uniqueness* of the risk minimizer and related conditions. In the next section, we develop new nonparametric tools to address this challenge.

4.2 Uniquness of risk minimizer

Our earlier Theorem 1 suggests the estimation of $\mathrm{sgn}(\Theta - \pi)$ via minimizing weighted classification loss. In order to establish the recovery guarantee from risk optimization, we shall address the uniqueness of the risk minimizer in the quotient space. The local behavior of Θ around π turns out to play 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., a histogram with bin width $2\Delta s$. Specifically, let $G(\pi)$ denote the cumulative distribution function (CDF) of $\Theta(\omega)$ under $\omega \sim \Pi$; i.e.,

$$G(\pi) := \mathbb{P}_{\omega \sim \Pi}[\Theta(\omega) \leq \pi].$$

We partition the range $[-1, 1] = \mathcal{N}^c \cup \mathcal{N}$, where the set $\mathcal{N}^c$ consists of *regular* π's for which the pseudo density is bounded; i.e,

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

and $\mathcal{N}$ otherwise. Let $|\mathcal{N}|_{\text{jump}}$ be the covering number of $\mathcal{N}$ with $2\Delta s$-bin's; i.e, $|\mathcal{N}|_{\text{jump}} = \lceil \mathrm{Leb}(\mathcal{N})/2\Delta s \rceil$, where $\mathrm{Leb}(\cdot)$ is the Lebesgue measure and $\lceil \cdot \rceil$ is the ceiling function. Intuitively, $\mathcal{N}$ consists of *irregular* π's for which the pseudo density is unbounded, and $|\mathcal{N}|_{\text{jump}}$ counts the *non-negligible jump points* while accounting for tolerance.

We introduce a notion of smoothness for the signal tensor. Our definition accounts for the discrete nature of $G(\cdot)$ via the introduction of tolerance Δs.

Definition 3 (α-smoothness of discrete distribution)**.** Fix $\pi \in \mathcal{N}^c$. A tensor Θ is called α-locally smooth at π, if there exist two constants $\alpha = \alpha(\pi) \geq 0$, $c = c(\pi) > 0$, independent of tensor dimension, such that

$$\sup_{\Delta s \leq t < \rho(\pi, \mathcal{N})} \frac{G(\pi + t) - G(\pi - t)}{t^\alpha} \leq c, \quad \text{with} \quad \rho(\pi, \mathcal{N}) := \min_{\pi' \in \mathcal{N}} |\pi - \pi'| + \Delta s. \tag{10}$$

Here $\rho(\pi, \mathcal{N})$ denotes the adjusted distance from π to the nearest point in $\mathcal{N}$. We make the convention that $\rho(\pi, \mathcal{N}) = 2$ (which equals the range of $\pi \in [-1, 1]$) when $\mathcal{N} = \emptyset$, and

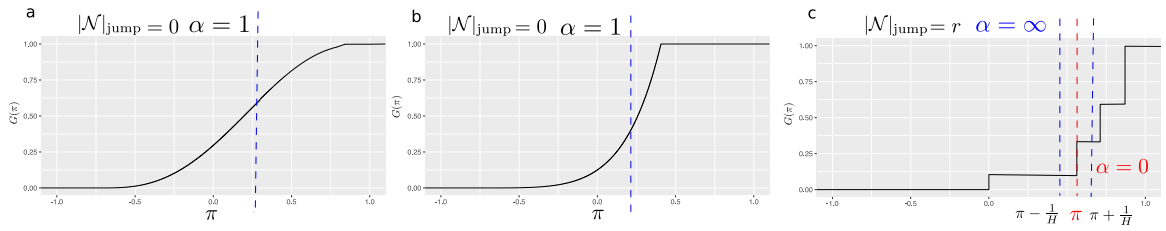


Figure 2: Three examples of CDF with smoothness index α at π depicted in dashed lines. Figure (a)-(b): Function $G(\pi)$ has $\alpha = 1$ with $\mathcal{N} = \emptyset$, because the $G(\pi)$ has finite pseudo density in the range of π. Figure (c): Function $G(\pi)$ has $\alpha = \infty$ at most π's (in blue), except for a total of $|\mathcal{N}|_{\text{jump}} = r$ non-negligible jump points (in red).

$\alpha = \infty$ when the numerator in (10) is zero. A tensor Θ is called α-globally smooth, if (10) holds with global constants $\alpha \geq 0$, $c > 0$ for all $\pi \in \mathcal{N}^c$.

Remark 4 (Interpretation of smoothness)**.** Figure 2 illustrates three examples of the CDF with various $(\alpha, \mathcal{N})$. Here, the local $\alpha \geq 0$ quantifies the growth speed in G at level π, and $|\mathcal{N}|_{\text{jump}} \geq 0$ counts the non-negligible jump points, both accounting for a small tolerance Δs. The value of α depends on both the sampling distribution $\omega \sim \Pi$ and the behavior of $\Theta(\omega)$. A small value of $\alpha = 0$ indicates heavy mass concentration at level π, or equivalently, when $G(\pi)$ jumps at π; a large value of $\alpha = \infty$ indicates nearly no concentration at level π, or equivalently, when $G(\pi)$ remains flat; an intermediate case is $\alpha = 1$ when $G(\pi)$ has a finite non-zero pseudo density in the vicinity of π.

We show that the α-smoothness with $\alpha \neq 0$ implies the essential uniqueness for the risk minimizer in (9). For two tensors Θ_1, Θ_2, define the mean absolute error (MAE) by

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

We now reach the main theorem in this section.

Theorem 2 (Perturbation bound of risk minimizer)**.** Assume $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ is α-globally smooth. Then, for all $\pi \in \mathcal{N}^c$ and all $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, we have

$$\text{MAE}(\text{sgn}\mathcal{Z},\ \text{sgn}(\Theta - \pi)) \lesssim C(\pi)\,[\text{Risk}(\mathcal{Z}) - \text{Risk}(\Theta - \pi)]^{\alpha/(\alpha+1)} + \Delta s, \tag{11}$$

where $C(\pi) > 0$ is a constant independent of $\mathcal{Z}$; the specific form of $C(\pi)$ is provided in Appendix C. In particular, when $\alpha \neq 0$, the risk minimizer is *uniquely* equal to $(\Theta - \pi)$, up to sign equivalence and a small tolerance Δs.

Remark 5 (Uniqueness and smoothness)**.** The bound (11) bounds the deviation in $\text{MAE}(\cdot, \cdot)$ by the difference in $\text{Risk}(\cdot)$. We conclude that the condition $\alpha \neq 0$ ensures the essential uniqueness of risk minimizer. Moreover, our result establishes the general recovery stability of sign tensor $\text{sgn}(\Theta - \pi)$ from weighted classification. The smoothness index α controls the worst-case perturbation around the risk minimizer $\text{sgn}(\Theta - \pi)$ with respect to the function $\text{Risk}(\cdot)$. We find that a higher value of α implies more stable recovery. This observation is consistent with intuition, because the best case $\alpha = \infty$ corresponds to easier estimation of $\text{sgn}(\Theta - \pi)$ with no point mass around decision boundary, whereas the worst case $\alpha = 0$ corresponds to hard estimation with a heavy mass around decision boundary.

Remark 6 (Comparison with hypergraphon models)**.** Our notion of smoothness is related to, but also distinctive from, nonparametric hypergraphon models (Xu, 2018; Balasubramanian, 2021). In both approaches, the signal tensor is interpreted as a multivariate function from domain space $\mathbb{R}^{d_1\times\cdots\times d_K}$ to range space $[-1,1]$. Here, we consider an α-smooth cumulative distribution function $G\colon [-1,1]\to\mathbb{R}$ in the range space; this is in contrast to nonparametric (hyper)graphon models that consider an α-smooth multivariate function $f\colon \mathbb{R}^{d_1\times\cdots\times d_K}\to\mathbb{R}$ in the domain space. The localness in our approach is determined by the range space $[-1,1]$, whereas the localness in (hyper)graphon model is determined by the domain space $[d_1]\times\cdots\times[d_K]$. The benefit bears the analogy of Lebesgue vs. Riemann integrals in functional analysis. The former is more appealing for tensor analysis, because the range space $[-1,1]$ is a simple scalar variable, whereas the domain space $[d_1]\times\cdots\times[d_K]$ is huge and multidimensional.

4.3 Why not l_q weight for $q\neq 1$?

The weighted classification loss in (7) penalizes the sign mismatches by the magnitudes of deviation from level π. One may ask whether same conclusion holds by penalizing more (or less) on the magnitude of deviation. Specifically, we examine the generalized l_q weighted classification risk defined by

$$\begin{aligned}\mathrm{Risk}_q(\mathcal{Z}) &= \mathbb{E}_{\omega\sim\Pi}\mathbb{E}_{\mathcal{Y}(\omega)}L_q(\mathcal{Z},\bar{\mathcal{Y}}_\Omega),\\ \text{where}\quad L_q(\mathcal{Z},\bar{\mathcal{Y}}_\Omega) &= \frac{1}{|\Omega|}\sum_{\omega\in\Omega}|\bar{\mathcal{Y}}(\omega)|^q|\mathrm{sgn}(\mathcal{Z}(\omega))-\mathrm{sgn}(\bar{\mathcal{Y}}(\omega))|.\end{aligned}$$

Here $q\geq 0$ is an integer controlling the penalization of magnitude $|\bar{\mathcal{Y}}(\omega)|$. We extend the our Theorem 1 from l_1 weights to l_q weights as follows.

Theorem 3 (Generalized weighted classification risk minimizer)**.** Consider the same set-up as in Theorem 1. Then, we have[2]

$$\operatorname*{arg\,min}_{\mathcal{Z}\in\mathbb{R}^{d_1\times\cdots\times d_K}}\mathrm{Risk}_q(\mathcal{Z}) = \begin{cases}\mathbb{E}(\mathcal{Y}-\pi)^q, & \text{if } q \text{ is an odd number},\\ \mathbb{E}\left[\mathcal{Y}-\pi\right]^q\mathbb{1}\{\mathcal{Y}\geq\pi\}-\mathbb{E}\left[\mathcal{Y}-\pi\right]^q\mathbb{1}\{\mathcal{Y}<\pi\}, & \text{if } q \text{ number}.\end{cases} \tag{12}$$

Furthermore, suppose either of the two conditions hold: (i) $p=1$, or (ii) $p\geq 0$ is a general integer, and the noise distribution $\mathcal{E}(\omega)$ is symmetric around origin. Then, $(\Theta-\pi)$ minimizes the l_q weighted classification risk; that is,

$$\operatorname*{arg\,min}_{\mathcal{Z}\in\mathbb{R}^{d_1\times\cdots\times d_K}}\mathrm{Risk}_q(\mathcal{Z}) = (\Theta-\pi). \tag{13}$$

By convention, all expressions above should be interpreted in an entrywise fashion. Performance bounds similar to Theorem 2 can be also derived for the generalized weighted classification risk minimizer.

2. The expectation here is taken with respect to $\mathcal{Y}(\omega)$ for each ω. For notational simplicity, the notion "=" in (12) and (13) should be interpreted as "one of the risk minimizers"(see Remark 3).

Remark 7 (Relationship between weights and noises)**.** Theorem 3 implies that, in the absence of symmetric noise, the generalized risk minimizer fails to recover $(\Theta - \pi)$ unless $p = 1$. For example, the risk minimizer becomes $(\text{Median}(\mathcal{E}) + \Theta - \pi)$ when $p = 0$ (i.e., the usual *unweighted* classification risk). The solution fails when $\text{Median}(\mathcal{E}) \neq 0$ in the absence of symmetry. The symmetric noise assumption is restrictive and often violated in practice. In particular, the distributions of Bernoulli, Binomial, Poisson residuals are intrinsically non-symmetric. Therefore, we choose l_1 weighted classification loss in (7).

5. Finite sample accuracy for signal tensor estimation

In previous sections we have established the sign series representation and its relationship to weighted classification. In this section, we present the estimation of $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ from sign tensor series. The main crux of our approach is a learning-reduction paradigm that connects a difficult $\mathbb{R}$-valued parameter estimation problem to a series of relatively simpler $\{0, 1\}$-valued weighted classification problems. We develop the algorithm, provide the estimation error bound, and apply the results to various scenarios.

5.1 Learning-reduction framework to signal estimation

Before describing our learning-reduction framework, we provide the intuition behind our method. In the two examples in Section 1.1, 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 Figure 1a-b 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 Figure 1c has the same sign pattern as a rank-2 tensor, since $\text{sgn}(\Theta(i, j, k) - \pi) = \text{sgn}(\max(i, j, k) - d(e^{\pi} - 1))$ (see Example 5 in Section 2.3).

The above observation suggests a general framework to estimate sign-representable signal tensors. Figure 3 illustrates the main steps of our algorithm. We propose to estimate the signal tensor $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ by taking the average over structured sign tensors

$$\hat{\Theta} = \frac{1}{2H + 1} \sum_{\pi \in \mathcal{H}} \text{sgn}\hat{\mathcal{Z}}_{\pi}, \quad \text{with} \quad \hat{\mathcal{Z}}_{\pi} = \underset{\mathcal{Z}:\, \text{rank}\mathcal{Z} \leq r}{\arg\min}\, L(\mathcal{Z}, \mathcal{Y}_{\Omega} - \pi), \tag{14}$$

where $\mathcal{H} = \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$ is the series of levels to consider, $H \in \mathbb{N}_+$ is a resolution parameter to be specified later (see Theorem 4), $L(\cdot, \cdot)$ is the ℓ_1 weighted classification loss in (7), and the rank constraint on $\mathcal{Z}$ follows from Theorem 1.

Our approach is built on the learning-reduction paradigm. 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 *polynomial* number of well-studied base problems suffice to recover Θ under the considered model (see the optimal choice of H in Theorem 4). The method therefore enjoys both statistical effectiveness and computational efficiency.

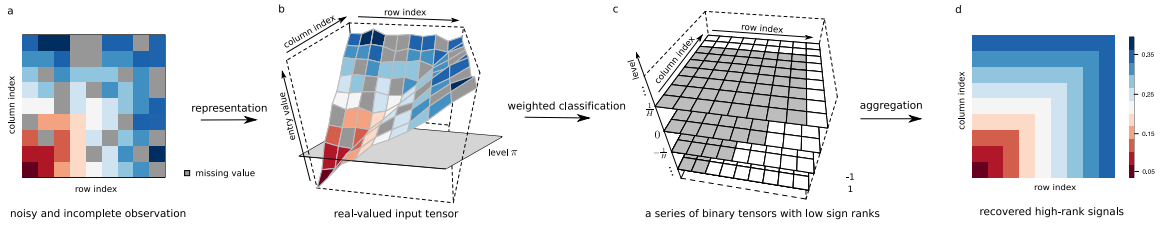


Figure 3: Illustration of our method in the context of an order-2 tensor (i.e. matrix). (a): a noisy, incomplete tensor input. (b)-(c): Estimation of sign tensor series $\mathrm{sgn}(\Theta - \pi)$ for $\pi \in \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$. (d): recovered signal $\hat{\Theta}$. The depicted signal is a full-rank matrix based on Example 5 in Section 2.3.

5.2 Implementation details

This section addresses the practical implementation of our estimation (14). We take a divide-and-conquer approach by dividing the full procedure into a meta algorithm and $(2H+1)$ base algorithms. The meta algorithm takes the average of $(2H+1)$ sign tensors, whereas each base algorithm estimates the tensor $\mathrm{sgn}(\Theta - \pi)$ given binary input $\mathrm{sgn}(\mathcal{Y}_\Omega - \pi)$. The full procedure is described in Algorithm 1.

Algorithm 1 Nonparametric tensor completion via learning-reduction

Input: A noisy and possibly 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: Base algorithm: Perform existing 1-bit tensor estimation algorithm (Ghadermarzy et al., 2018; Wang and Li, 2020; Hong et al., 2020) on $(\mathcal{Y}_\Omega - \pi)$ and obtain

$$\hat{\mathcal{Z}}_\pi = \underset{\mathcal{Z}:\mathrm{rank}(\mathcal{Z}) \le r}{\arg\min} \; L(\mathcal{Z}, \mathcal{Y}_\Omega - \pi). \tag{15}$$

3: **end for**

4: Meta algorithm: Average over estimated sign tensors

$$\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}(\hat{\mathcal{Z}}_\pi). \tag{16}$$

Output: Estimated signal tensor $\hat{\Theta}$.

The base algorithm (15) reduces to a well-studied low-rank 1-bit tensor estimation problem. A number of efficient algorithms with convergence guarantees are readily available for this problem (Ghadermarzy et al., 2018; Wang and Li, 2020; Hong et al., 2020; Han et al., 2022). For example, Wang and Li (2020) developed an alternating optimization algorithm, and Han et al. (2022) developed a projected gradient descent algorithm. For self-containedness, we summarize the alternating optimization algorithm here. More details can be found in previous work (Wang and Li, 2020; Hong et al., 2020). 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 $\boldsymbol{A}_k = [\boldsymbol{a}_1^{(k)}, \ldots, \boldsymbol{a}_r^k] \in \mathbb{R}^{d_k \times k}$, where tensor singular values are collected into the last factor $\boldsymbol{A}_K$. We numerically solve (15)

by optimizing one factor $\boldsymbol{A}_k$ at a time while holding others fixed. Each suboptimization reduces to a simple classification problem with classical vector-based decision variable. Following common practice in tensor optimization, we run the optimization from multiple initializations to locate a final estimate with the lowest objective value.

We emphasize that we did not attempt to propose a new base algorithm for (15) and its algorithmic convergence; indeed, both are not new in the literature. Our major contribution is a learning-reduction framework by adopting existing algorithms for a more challenging high-rank problems. We will show this approach achieves statistical-computational efficiency almost for free, i.e., at almost no extra statistical cost and only an extra poly(d) computational cost (see Theorem 4). The developed sign-representable tensor model unifies low-rank and high-rank tensors, thereby empowering exiting algorithms for broader implications.

In theory, extra technical assumptions are needed for the algorithmic convergence of the base algorithm (Wang and Li, 2020; Han et al., 2022). We omit the details here but instead assume the estimate (15) is attainable. We also note the existence of several 1-bit tensor algorithms that use surrogate loss (Genzel and Stollenwerk, 2020; He et al., 2017) instead of the loss in (14). Under certain regularity conditions, these algorithms can also be adopted in our setting. More details on the implementation will be discussed in Section 6.4.

Remark 8 (Distinction between our algorithm and classical 1-bit tensor algorithm)**.** The key novelty in our algorithm compared to classical 1-bit tensor algorithm lies in the aggregation step (16). Classical tensor algorithm takes $\hat{\mathcal{Z}}_\pi$ as output. By contrast, we dichotomize the estimate $\hat{\mathcal{Z}}_\pi$ by ignoring its magnitude, and then aggregate the sign series $\mathrm{sgn}(\hat{\mathcal{Z}}_\pi)$ into a new continuous-valued estimator $\hat{\Theta}$. Somewhat surprisingly, we find that the dichotomization not only preserves all information in the original signals, but also brings benefits of accuracy and flexibility over classical parametric methods.

5.3 Finite sample accuracy

Given a noisy incomplete tensor observation $\mathcal{Y}_\Omega$ from model (3), we cast the problem of estimating $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$ into a series of weighted classifications. We are particularly interested in the high-dimensional region, where $\min_{k\in[K]} d_k \to \infty$ while holding $r \lesssim \mathcal{O}(1)$ as fixed. For a cleaner exposition, we assume the loss (7) is bounded by a universal constant; the extension to unbounded loss is provided in Section 6. The next theorem establishes the finite sample accuracy for the estimates from Algorithm 1.

Theorem 4 (Estimation error for sign representable tensors)**.** Consider the model (3) with $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$. Suppose $\Theta(\omega)$ is an α-globally smooth tensor with non-negligible jump points collected in set $\mathcal{N}$. Let $\hat{\mathcal{Z}}_\pi$ and $\hat{\Theta}$ denote the estimates in (15) and (16) from Algorithm 1, respectively. Denote

$$d_{\max} = \max_{k\in[K]} d_k, \quad \text{and} \quad t_d = \frac{d_{\max} r \log|\Omega|}{|\Omega|}.$$

Under the bounded loss assumption, we have the following error bounds with a high probability at least $1 - \exp(-d_{\max} t_d)$.

(a) (Sign tensor estimation). For all $\pi \in \mathcal{N}^c$,

$$\mathrm{MAE}(\mathrm{sgn}(\hat{\mathcal{Z}}_\pi),\ \mathrm{sgn}(\Theta - \pi)) \lesssim t_d^{\frac{\alpha}{\alpha+2}} + \frac{1}{\rho^2(\pi, \mathcal{N})} t_d. \tag{17}$$

(b) (Signal tensor estimation). For any resolution parameter $H \in \mathbb{N}_+$,

$$\mathrm{MAE}(\hat{\Theta},\ \Theta) \lesssim \underbrace{(t_d \log H)^{\frac{\alpha}{\alpha+2}}}_{\text{error inherited from sign estimation}} + \underbrace{\frac{1 + |\mathcal{N}|_{\mathrm{jump}}}{H}}_{\text{bias}} + \underbrace{t_d (H \log H)}_{\text{variance}}. \tag{18}$$

In particular, setting $H = \sqrt{(1+|\mathcal{N}|_{\mathrm{jump}})/t_d} \lesssim \mathrm{poly}(d_{\max})$ yields the tightest upper bound in (18).

In the special case of full observation with equal dimension $d_1 = \cdots = d_K = d$ and bounded $|\mathcal{N}|_{\mathrm{jump}} \lesssim \mathcal{O}(1)$, our signal tensor estimate achieves convergence

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim r d^{-(K-1)\min(\frac{\alpha}{\alpha+2},\ \frac{1}{2})} \log^2 d, \quad \text{by setting } H \asymp d^{(K-1)/2}. \tag{19}$$

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 Proposition 1). Our result also reveals that the convergence becomes favorable as the order of data tensor increases.

Remark 9 (Sign estimation compared to existing work)**.** The bound (17) is a finite-sample version of Theorem 2. The result demonstrates the polynomial decay of sign errors with sample size $|\Omega|$. Our error bound (17) improves the existing work on weighted classification. Existing work (Tsybakov, 2004; Xu et al., 2020) considered only a finite number of π's, and provided only the first term in the bound (17). By contrast, our bound quantifies the full dependence on the level π and establishes the recovery of $\mathrm{sgn}(\Theta - \pi)$ *uniformly* over all possible $\pi \in \mathcal{N}^c$. It turns out both terms in the bound (17) are crucially in the signal error bound in (18): the first term contributes to the inherited sign estimation error, whereas the second term contributes to the variance term in sign averaging.

Remark 10 (Signal estimation)**.** The bound (18) demonstrates our main results for non-parametric signal estimation. We make three remarks.

• (Error decomposition) The bound (18) reveals three sources of errors: the estimation error for sign tensors, the bias from sign series representations, and the variance thereof. The resolution parameter $H \in \mathbb{N}_+$ determine the number of sign tensors to average in the estimation. The best choice of $H = \sqrt{(1 + |\mathcal{N}|_{\mathrm{jump}})/t_d}$ balances the bias-variance tradeoff. A larger value of H reduces the approximation bias but renders the sign estimation harder near mass points, and vice versa.

• (From sign to signal estimation) We find that the signal estimation error (18) is generally no better than the corresponding sign error (17). In the special case as in (19) with $\alpha = \infty$, the sign estimation (17) reaches a fast rate $\mathcal{O}(d^{-(K-1)})$ whereas the signal tensor estimation reaches a slow rate $\mathcal{O}(d^{-(K-1)/2})$. The phenomenon is to be expected, since magnitude estimation is harder than sign estimation.

• (Robustness) We also find that the signal estimation is robust to finitely many off-target sign estimations, as long as the majorities are accurate. This can be seen by the order equivalence of $|\mathcal{N}|_{\text{jump}} = 0$ vs. $|\mathcal{N}|_{\text{jump}} \asymp \mathcal{O}(1)$ in the bound (18). Recall that $|\mathcal{N}|_{\text{jump}}$ counts the non-negligible jump points for which sign tensors are nonrecoverable from classification risks (see red line in Figure 2c). Nevertheless, the signal tensor is still estimable, because the nearby sign estimations (blue lines in Figure 2c) provide the magnitude information. The fact shows the benefit of sign aggregation approach to signal estimation.

Remark 11 (Technical novelty in the proof)**.** The main challenge in the proof of Theorem 4 is that our estimates are *not* likelihood-based. In particular, our loss function is not restricted to a particular noise distribution; instead only weak first-order moment assumption is invoked. Such flexibility renders classical likelihood-based tensor analysis (Wang and Li, 2020; Ghadermarzy et al., 2018; Zhang and Xia, 2018) non-applicable in our setting. We develop new empirical process tools to address this challenge.

The proof of Theorem 4 consists of three main ingredients. We first leverage the α-smoothness to provide a sharp classification error faster than the usual root-n convergence. The improvement stems from a variance-to-mean relationship in classification loss and a careful *local* analysis of empirical process (see Appendix D). The result implies that, in the local region of risk minimizer $\text{sgn}(\Theta - \pi)$, the estimate $\hat{\mathcal{Z}}_\pi$ converges more quickly than the simple uniform convergence results would suggest. The second step is to convert the risk error into the mean absolute error by Theorem 2. The last step is to aggregate the sign errors into the continuous-valued signal estimation error. A careful error analysis reveals the joint contribution from both sign aggregations and variance-bias trade-off.

Finally, we apply our Theorem 4 to the problem of tensor completion. The following corollary reveals the sample complexity for nonparamtric tensor completion.

Corollary 1 (Sample complexity for nonparametric completion)**.** Assume the same conditions of Theorem 4 and bounded $|\mathcal{N}|_{\text{jump}}$. Then, with a high probability at least $1 - \exp(-d_{\max} t_d)$,

$$\text{MAE}(\hat{\Theta}, \Theta) \to 0, \quad \text{as} \quad \frac{|\Omega|}{d_{\max} r \log^2 |\Omega|} \to \infty.$$

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 ($d_{\max} r$) 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.

5.4 Revisiting earlier examples

We apply our method to the main examples in Section 2.3. For simplicity, suppose $\omega \sim \Pi$ is the uniform sampling. The comparison with existing literature is summarized in Table 2. Numerical comparisons are provided in Section 7.

Example 2 (TBMs)**.** Consider a tensor block model with r multiway blocks. The tensor block model $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ is ∞-globally smooth. The number of non-negligible jump points

Model	α	$\lvert\mathcal{N}\rvert_{\text{jump}}$	Our rate (power of d)	Comparison with previous results
Sign representable tensor	≥ 0	Finite	$-(K-2)\min(\frac{\alpha}{\alpha+2}, \frac{1}{2})$	New.
Tensor block model	∞	Finite	$-(K-1)/2$	Minimax optimal as in (Wang and Zeng, 2019)
Single index model	1	0	$-(K-1)/3$	New for general $K > 3$; Improves the previous rate -1/4 for $K = 2$ (Ganti et al., 2015).
Generalized linear model	1	0	$-(K-1)/3$	Close to minimax rate (Wang and Li, 2020).
Structure with repeating entries	∞	d	$-(K-2)/2$	New.

Table 2: Summary of our statistical rates compared to existing works under different models. For notational simplicity, we present error rates assuming equal tensor dimension in all modes and finite $|\mathcal{N}|_{\text{jump}}$ for the smooth tensor model. Here $K \geq 2$ denotes the tensor order and d denotes the tensor dimension.

$|\mathcal{N}|_{\text{jump}}$ equals the number of distinct values in Θ. The CDF G satisfies $\alpha = \infty$ for all regular π's, since no point mass at $\pi \in \mathcal{N}^c$ (see Figure 2c for an example). Applying our Theorem 4 implies a rate $\tilde{\mathcal{O}}(d^{-(K-1)/2})$ by taking $\alpha = \infty$ and $|\mathcal{N}|_{\text{jump}} \leq r \lesssim \mathcal{O}(1)$. This rate agrees with the previous root-mean-square error (RMSE) for block tensor estimation (Wang and Zeng, 2019). The rate is known to be minmax optimal, thereby suggesting the sharpness of our bound in this example.

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 Θ has bounded pseudo density with $\alpha = 1$ (see Figure 2b-c for an example). Applying our Theorem 4 yields the estimation error $\tilde{\mathcal{O}}(d^{-(K-1)/3})$. 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 high-rank tensors. The nonparametric rate is the price one has to pay for not knowing the form $g(\cdot)$ 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). Recent work (Xu, 2018) establishes the RMSE rate $\tilde{\mathcal{O}}(d^{-1/3})$ for Lipschitz bivariate graphon models for matrices. The rate is a special case of our Theorem 4 by setting tensor order $K = 2$, thereby suggesting the sharpness of our error bound.

Example 5 (Structured tensors with repeating entries)**.** Consider the high-rank order-K (d, d, d)-dimensional tensor with entries $\Theta(i_1, \cdots, i_K) = \log(1 + \max_{k\in[K]} i_k/d)$. We have known $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ from Section 2.3. Now we conclude Θ is ∞-globally smooth. The number of non-negligible jump points equals d, because $|\mathcal{N}|_{\text{jump}} = \text{Card}\{\log(1 + i/d) : i \in [d]\} = d$, where $\text{Card}\{\cdot\}$ denotes the cardinality of the set. The CDF G satisfies $\alpha = \infty$ for all $\pi \in \mathcal{N}^c$. Applying our Theorem 4 with $\alpha = \infty$ and $|\mathcal{N}|_{\text{jump}} = d$ yields the rate $\tilde{\mathcal{O}}(d^{-(K-2)/2})$.

6. Extensions

Recall our signal plus noise model in (3),

$$\mathcal{Y} = \Theta + \mathcal{E}, \quad \mathbb{E}(\mathcal{E}) = 0,$$

where $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ is an α-smooth sign representable signal tensor of interest, and $\mathcal{E}$ consists of zero-mean, independent noise entries. In our main theory of earlier sections 4-5, we have imposed rather weak assumptions on the noise, in that only three aspects are invoked: (i) entrywise independence of $\mathcal{E}$ (conditional on Θ), (ii) zero-mean constraint $\mathbb{E}(\mathcal{E}) = 0$; and (iii) boundedness of weighted loss (for finite sample accuracy results only). Other aspects of the noise (e.g. variance, symmetry, distribution) are irrelevant. In this section, we specialize our results to various problems and provide extension to unbounded weighted loss.

6.1 Sub-Gaussian tensor denoising

Sub-Gaussian tensor denoising is a popular problem in literature. A range of parametric structures have been considered in earlier work, including CP-low rankness (Anandkumar et al., 2014), sparsity (Allen, 2012), and blockness (Wang and Zeng, 2019). Here, we generalize our nonparametric sign representable model (3) to sub-Gaussian tensor denoising problem. Specifically, suppose the noise tensor $\mathcal{E}$ in (3) consists of zero-mean, independent sub-Gaussian entries with variance bounded by σ^2.

Assumption 1 (Sub-Gaussian noise)**.** The noise entries $\mathcal{E}(\omega)$ are independent zero-mean sub-Gaussian random variables with variance proxy $\sigma^2 > 0$; i.e,

$$\mathbb{P}(|\mathcal{E}(\omega)| \geq B) \leq 2e^{-B^2/2\sigma^2}, \text{ for all } B > 0.$$

The sub-Gaussian noise satisfies the aforementioned noise assumptions (i)-(ii) but not (iii). In particular, the classification loss (7) becomes unbounded due to the presence of weights $|\bar{\mathcal{Y}}(\omega)|$. The following Corollary provides our nonparametric estimation error for sub-Gaussian tensor denoising.

Corollary 2 (sub-Gaussian tensor denoising with nonparametric signals)**.** Consider the same setup as in Theorem 4. Suppose that Assumption 1 is used in place of the bounded loss assumption. For simplicity, assume $|\mathcal{N}|_{\text{jump}} \lesssim \mathcal{O}(1)$ and $d_1 = \cdots = d_K = d$. Set

$$t_d = \frac{\sigma^2 dr \log d \log |\Omega|}{|\Omega|} \quad \text{and} \quad H \asymp \sqrt{\frac{1}{t_d}}.$$

Then, with a high probability at least $1 - \exp(-dt_d)$,

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim t_d^{\min\left(\frac{\alpha}{\alpha+2}, \frac{1}{2}\right)}.$$

We find that the (unbounded) sub-Gaussian noise incurs only an additional $\log d$ factor compared to Theorem 4.

6.2 Probabilistic tensor estimation

Binary tensor problems commonly arise in recommendation system and network analysis. Examples include context-based recommendation system (Adomavicius and Tuzhilin, 2011), multi-relational social networks (Nickel et al., 2011), and brain connectivity network (Wang et al., 2019). Here, we generalize our nonparametric model to binary tensors.

Specifically, assume that we observe a binary tensor $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$ whose entries are realizations of independent Bernoulli random variables, such that,

$$\mathcal{Y}(\omega) \sim \text{Bernoulli}\left(\Theta(\omega)\right), \text{ for all } \omega \in \Omega. \tag{20}$$

Here the signal tensor, $\Theta \in \mathscr{P}_{\text{sgn}}(r)$, represents the probability tensor of interest. Notice that model (20) can be represented as the additive model (3), by setting the noise as the Bernoulli residual $\mathcal{E} = \mathcal{Y} - \Theta$. Since the weighted loss (7) is bounded in this setting, directly applying Theorems 4 yields the estimation error.

Corollary 3 (Probabilistic tensor estimation with nonparametric signals)**.** Consider the same setup as in Theorem 4. Suppose a binary tensor is observed from (20). For simplicity, assume $|\mathcal{N}|_{\text{jump}} \lesssim \mathcal{O}(1)$ and $d_1 = \cdots = d_K = d$. Set

$$t_d = \frac{dr \log |\Omega|}{|\Omega|} \quad \text{and} \quad H \asymp \sqrt{\frac{1}{t_d}}.$$

Then, with a high probability at least $1 - \exp(-dt_d)$,

$$\text{MAE}(\hat{\Theta}, \Theta) \lesssim t_d^{\min\left(\frac{\alpha}{\alpha+2}, \frac{1}{2}\right)}.$$

We find that the entries of nonparametric tensor estimator $\hat{\Theta}$ automatically fall into the valid probability range $[0,1]$. This is because in the aggregation step (16), the sign tensors $\text{sgn}(\hat{\mathcal{Z}}_\pi) = \mathbf{1} \otimes \cdots \otimes \mathbf{1}$ are constant tensors for a total of $(H+1)$ non-positive levels $\pi \leq 0$. Therefore, $\hat{\Theta}(\omega) \in [\frac{(H+1)-H}{2H+1}, \frac{(H+1)+H}{2H+1}] \subset [0,1]$.

6.3 Binomial tensor problem

Binomial tensor problems appear when the observations are success counts from a limited numbers of trails. For example, the human mortality databset (Jdanov et al., 2019) provides the death counts and the total numbers of individuals for each combination (country, age, year). This dataset is naturally summarized as a three-way tensor of country $\times$ age $\times$ year, where the entries can be modeled as independent binomial trials. Here, we generalize our nonparametric model to binomial tensors.

Specifically, suppose that we observe a success count tensor $\mathcal{Y} \in \mathbb{N}^{d_1 \times \cdots \times d_K}$ from a total trial count tensor $\mathcal{P} \in \mathbb{N}_+^{d_1 \times \cdots \times d_K}$, where $0 \leq \mathcal{Y} \leq \mathcal{P}$ in an entrywise fashion. Assume that $\mathcal{Y}$ consists of independent binomial counts given the total trial counts $\mathcal{P}$, such that,

$$\mathcal{Y}(\omega) \sim \text{Binomial}\left(\mathcal{P}(\omega), \Theta(\omega)\right), \quad \text{for all } \omega \in \Omega. \tag{21}$$

The goal is to estimate the success probability $\Theta \in \mathscr{P}_{\mathrm{sgn}}(r)$ from the observed tensor $\mathcal{Y}_\Omega$ given total trial counts $\mathcal{P}$. Then, we have the following theoretical guarantee for the binomial tensor model.

Corollary 4 (Binomial tensor problem with nonparametric signals)**.** Consider the same setup as in Theorem 4. Suppose a success count tensor is observed from (21). For simplicity, assume $|\mathcal{N}|_{\mathrm{jump}} \lesssim \mathcal{O}(1)$, $\|\mathcal{P}\|_{\max} \lesssim \mathcal{O}(1)$, and $d_1 = \cdots = d_K = d$. Set

$$t_d = \frac{dr \log |\Omega|}{|\Omega|} \quad \text{and} \quad H \asymp \sqrt{\frac{1}{t_d}}.$$

Then, with a high probability at least $1 - \exp(-dt_d)$,

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim \frac{1}{\min_{\omega \in [d_1] \times \cdots \times [d_K]} \mathcal{P}(\omega)} t_d^{\min\left(\frac{\alpha}{\alpha+2}, \frac{1}{2}\right)}.$$

Corollary 4 is a generalization of Corollary 3 from Bernoulli trails with $\mathcal{P} = \mathbf{1} \otimes \cdots \otimes \mathbf{1}$ to Binomial trials with general $\mathcal{P} \in \mathbb{N}_+^{d_1 \times \cdots \times d_K}$.

6.4 Extension to hinge loss

In earlier sections, we have established our main results under the weighted classification loss

$$L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \underbrace{|\bar{\mathcal{Y}}(\omega)|}_{l_1 \text{ weight}} \times \underbrace{|\mathrm{sgn}\mathcal{Z}(\omega) - \mathrm{sgn}\bar{\mathcal{Y}}(\omega)|}_{\text{classification loss}},$$

where $|\mathrm{sgn}(\mathcal{Z}) - \mathrm{sgn}(\bar{\mathcal{Y}})|$ is the canonical classification loss. Here, we generalize the canonical classification loss $\ell(z, y) = |\mathrm{sgn}z - \mathrm{sgn}y|$ to (unbounded) hinge loss $F(m) = (1-m)_+$. Hinge loss is commonly used as a surrogate classification loss due to its continuity (Bartlett et al., 2006; Genzel and Stollenwerk, 2020; He et al., 2017).

Specifically, we establish the parallel theory of Theorem 4 for hinge loss. The only difference in setup is that, the estimates $\mathcal{Z}_\pi$, $\hat{\Theta}$ in (15)-(16) are now replaced by

$$\hat{\mathcal{Z}}_{\pi,F} = \operatorname*{arg\,min}_{\mathrm{rank}(\mathcal{Z}) \le r} \sum_{\omega \in \Omega} |\bar{\mathcal{Y}}(\omega)| F(\mathcal{Z}(\omega)\mathrm{sgn}\bar{\mathcal{Y}}(\omega)) + \lambda_\pi \|\mathcal{Z}\|_F^2, \quad \hat{\Theta}_F = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}(\hat{\mathcal{Z}}_{\pi,F}) \tag{22}$$

where $F(\cdot) = (1-m)_+$ is the hinge loss, and $\lambda_\pi > 0$ is the penalty parameter. We use $\mathrm{Risk}_F(\cdot)$ to denote the *surrogate weighted classification loss*, defined similarly as in (8), with hinge loss used in place of canonical loss. A nice property of hinge loss is the following *Fisher consistency bound* (Scott, 2011),

$$\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\Theta - \pi) \lesssim \mathrm{Risk}_F(\mathcal{Z}) - \mathrm{Risk}_F(\Theta - \pi), \text{ for all } \pi \in [-1, 1] \text{ and all tensors } \mathcal{Z}. \tag{23}$$

The Fisher consistency enables us to relate the excess risk of the hinge loss to that of canonical loss.

Assumption 2 (Approximation bias)**.** Denote $n = |\Omega|$. For all $\pi \in \mathcal{N}^c$, assume there exist a sequence of tensors $\mathcal{Z}_\pi^{(n)} \in \mathscr{P}_{\text{sgn}}(r)$, such that $\text{Risk}_F(\mathcal{Z}_\pi^{(n)}) - \text{Risk}_F(\Theta - \pi) \le a_n$, for some sequence $a_n \to 0$ as $n \to \infty$. Furthermore, assume $\|\mathcal{Z}_\pi^{(n)}\|_F \le J$ for some constant $J > 0$.

Assumption 2 quantifies the representation capability of $\mathscr{P}_{\text{sgn}}(r)$ for the true $\text{sign}(\Theta - \pi)$. Under Assumption 2, the hinge estimate (22) enjoys statistical efficiency.

Theorem 5 (Hinge loss based estimation)**.** Consider the same setup as in Theorem 4. For simplicity, assume $|\mathcal{N}|_{\text{jump}} \lesssim \mathcal{O}(1)$ and $d_1 = \cdots = d_K = d$. Denote

$$n = |\Omega|, \quad t_n = \frac{rd \log n}{n}, \quad \lambda_\pi \asymp t_n^{\frac{\alpha+1}{\alpha+2}} + \frac{t_n}{\rho(\pi, \mathcal{N})}, \text{ and } H \asymp \sqrt{\frac{1}{t_n}}.$$

Suppose the Assumption 2 with $a_n \lesssim t_n^{(\alpha+1)/(\alpha+2)}$ holds in place of the bounded loss assumption. Then, with a high probability at least $1 - \exp(-nt_n)$,

$$\text{MAE}(\hat{\Theta}_F, \Theta) \lesssim t_n^{\min\left(\frac{\alpha}{\alpha+2}, \frac{1}{2}\right)}. \tag{24}$$

Remark 12 (Comparison between canonical loss and hinge loss)**.** The estimation under canonical loss (Theorem 4) requires no penalization with $\lambda_\pi = 0$, because only the sign, but not the magnitude, affects the weighted classification risk. One can impose norm constraint$\|\mathcal{Z}\|_F = 1$ in the empirical risk minimization without altering the solution. By contrast, the hinge loss is scale-sensitive, rending the possible unboundedness of $L(\cdot, \cdot)$. We therefore impose penalization to control the magnitude of the $\|\mathcal{Z}\|_F$ and the local complexity. We find that the resulting estimation enjoys the fast convergence under well tuned λ_π.

Other large-margin losses are also applicable, such as logistic loss $F(m) = \log(1+e^{-m})$ (Wang and Li, 2020) and ψ-loss $F(m) = \min(1, (1-m)_+)$ with $m_+ = \max(m, 0)$ (Shen et al., 2003). In principle, users can choose their own favorite large-margin losses. Similar theoretical accuracy for other large margin losses are possible provided that the chosen loss satisfies Fisher consistency (23) and Assumption 2. The comparison between various large-margin losses has been studied before (Bartlett et al., 2006).

7. Numerical experiment

7.1 Synthetic data

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 3. The simulation covers a wide range of complexity, including block tensors, transformed low rank tensors, structured tensors with repeating entries umder logarithm and exponential transformation. 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.

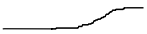
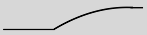
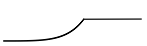
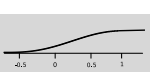
Simulation	Signal Tensor Θ	Rank	Sign Rank	α	$\lvert\mathcal{N}\rvert_{\text{jump}}$	CDF	Noise
1	$\mathcal{C} \times \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \boldsymbol{M}_3$	3^3	$\leq 3^3$	∞	$\leq 3^3$		Uniform $[-0.3, 0.3]$
2	$\lvert\boldsymbol{a} \otimes \mathbf{1} \otimes \mathbf{1} - \mathbf{1} \otimes \boldsymbol{a} \otimes \mathbf{1}\rvert$	d	≤ 3	1	0		Normal $N(0, 0.15)$
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 $N(0, 0.15)$

Table 3: 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, $\boldsymbol{a} = d^{-1}(1, 2, \ldots, d)^T \in \mathbb{R}^d$, $\mathcal{Z}_{\max}$ and $\mathcal{Z}_{\min}$ are order-3 tensors with entries $\max(i, j, k)/d$ and $\min(i, j, k)/d$, respectively.

Figure 4 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 5 shows the completion error against observation fraction. We fix $d = 40$ and gradually increase the observation fraction $|\Omega|/d^3$ 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.

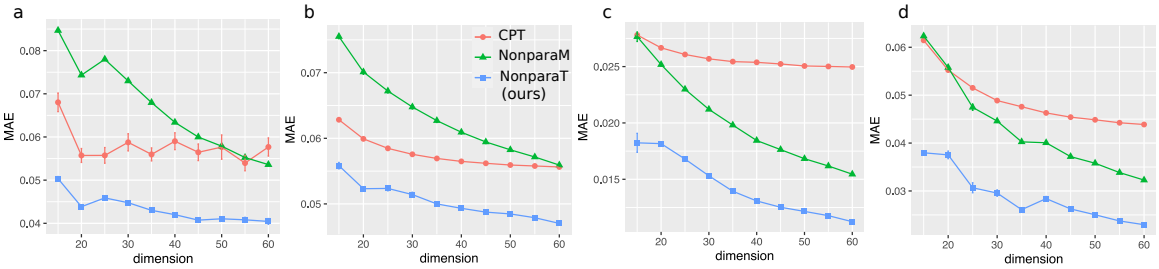


Figure 4: Estimation error versus tensor dimension. Panels (a)-(d) correspond to simulation models 1-4 in Table 3.

7.2 Brain connectivity analysis

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). The

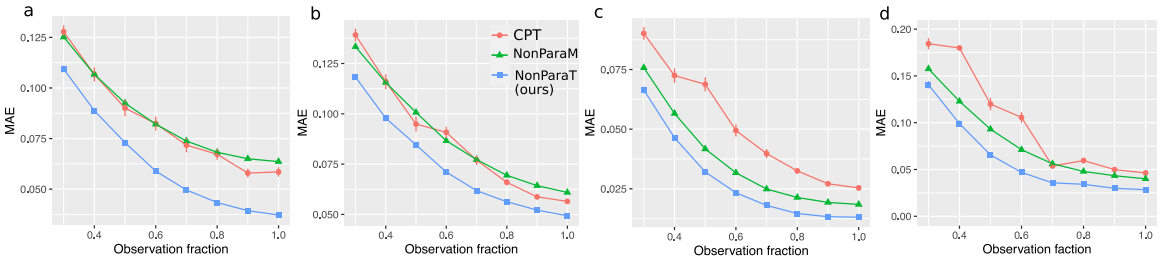


Figure 5: Completion error versus observation fraction. Panels (a)-(d) correspond to simulation models 1-4 in Table 3.

first data tensor consists of binary entries only, and the second data tensor consists of continuous-valued entries.

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.

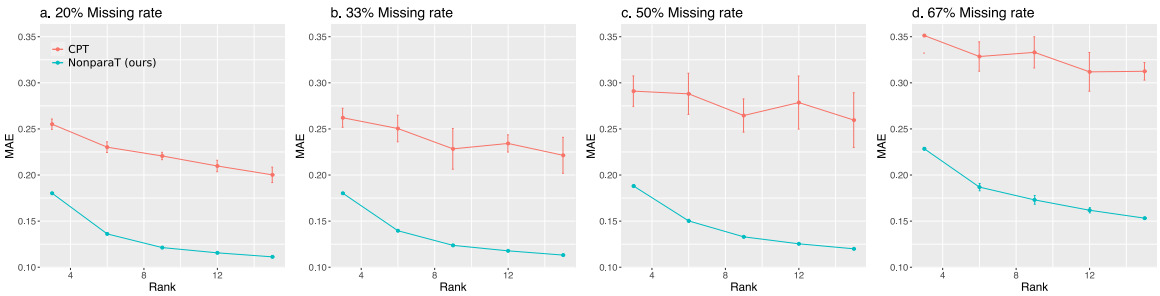


Figure 6: Estimation error versus rank under different missing rate. Panels (a)-(d) correspond to missing rate 20%, 33%, 50%, and 67%, respectively. Error bar represents the standard error over 5-fold cross-validations.

Figure 6 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 6 and Table 4, demonstrating the stability over CPT. One possible reason is that that our estimate is guaranteed to be in $[0, 1]$ (for probabilistic tensor estimation) whereas CPT estimation may fall outside the valid range $[0, 1]$.

We next investigate the pattern in the estimated signal tensor. Figure 7a 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 lobes. The result is consistent

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 4: 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 $H = 20$.

with recent research on brain connectivity with intelligence (Li et al., 2009; Wang et al., 2017).

7.3 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 4 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.

We next examine the estimated signal tensor $\hat{\Theta}$ from our method. Figure 7b 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.

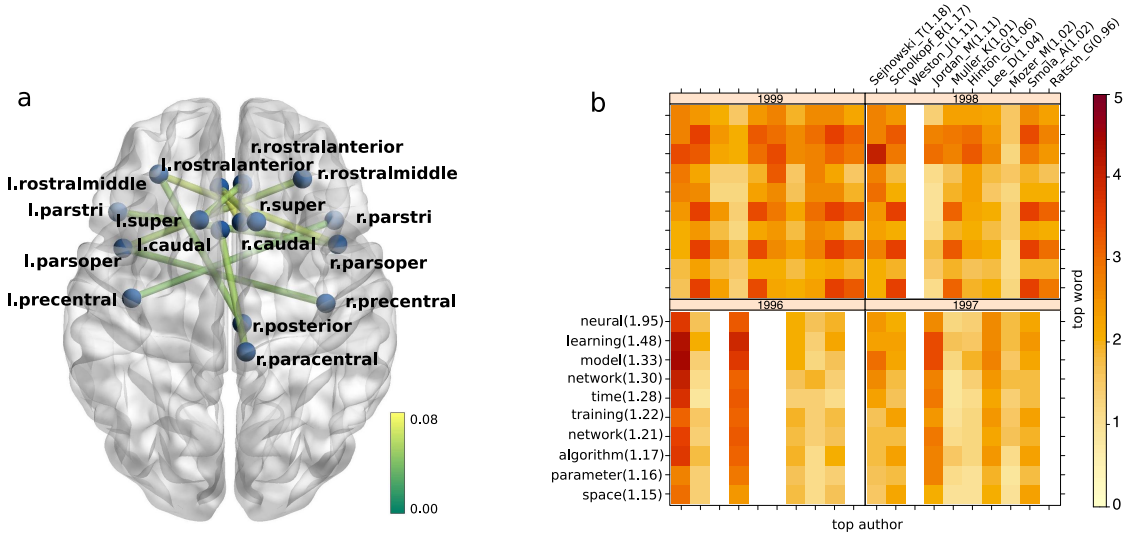


Figure 7: Estimated signal tensors in the data analysis. (a) top edges associated with IQ scores in the brain connectivity data. The color indicates the estimated IQ effect size. (b) top authors and words for years 1996-1999 in the NIPS data. Authors and words are ranked by marginal averages based on $\hat{\Theta}$, where the marginal average is denoted in the parentheses.

8. Discussion

We have developed a tensor estimation method that addresses both low- and high-rankness based on sign series representation. Our work provides a nonparametric framework for tensor estimation, and we establish accuracy guarantees for recovering a wide range of structured tensors. Our proposed learning-reduction strategy empowers existing algorithms for broader implication, thereby connecting the low-rank (parametric) tensors and high-rank (nonparametric) tensors. We hope the work opens up new inquiry that allows more researchers to contribute to this field.

There are several possible extensions from our work. In our theory, we assume the true sign-rank r is known or otherwise has been consistently estimated; the adaptivity to unknown r is empirically addressed by cross-validation. In principle, a larger r leads to a smaller approximation bias but a larger estimation variance; vice versa for a smaller r. When we set the rank smaller than true signal rank r, this would incur extra approximation bias. The level of approximation bias depends on the spectral complexity of the signal tensor. In the matrix case, the approximation bias is well quantified by tail eigenvalues; however, such extension to higher-order tensors is known to be very challenging, partly due to lack of spectral theory. We will leave the optimality of unknown rank as future research.

Exploring the minimax nonparametric rate among polynomial-time tensor algorithms warrants future work. In the special setting of tensor order $K = 2$ (i.e. matrix case) with smoothness $\alpha = 1$, our method achieves the error rate $\tilde{\mathcal{O}}(d^{-(K-1)/3})$. This rate agrees with RMSE rate obtained by Lipschitz graphon model (Xu, 2018). It has been conjectured that $\tilde{\mathcal{O}}(d^{-1/3})$ is the best computationally efficient rate for $K = 2$ (Xu, 2018; Zhang et al., 2022).

We conjecture the similar phenomenon extends to higher-order tensors ($K \geq 3$); a full exploration will be the future work.

Our model is related to but distinct from simple hypergraphons (Balasubramanian, 2021) as mentioned in Remark 6. The improvement of our approach over simple smooth hypergraphon bears the analogy of Lebesque vs. Riemann integrals in functional analysis. In addition, simple hypergraphon uses random design points, whereas our example uses deterministic design points. These two choices of designs lead to a notable different analysis in the same spirit as random- vs. fixed-designs in nonparametric regression. Whether it is possible to extend our theory to general hypergraphon is an interesting question for future research.

Finally, our learning-reduction algorithm uses evenly spaced grid of levels, $\mathcal{H} = \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$. Our theory continues to hold true for non-evenly spaced grids, as long as the spacings converge at a same rate. In principle, using non-evenly spaced grid may improve the local adaptivity, and we conjecture that the optimal grid design depends on the true (but unknown) shape of the CDF $G(\cdot)$. In real applications, however, we usually have no prior information on function smoothness of α. The lack of prior knowledge makes optimal grid design sensitive to model misspecification. Therefore, we choose to present the evenly spaced grid to achieve a good balance between theory and empirical applicability.

Acknowledgements

This research is supported in part by NSF CAREER DMS-2141865, NSF grant DMS-1915978, DMS-2023239 and Wisconsin Alumni Research Foundation.

References

Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.

Genevera Allen. Sparse higher-order principal components analysis. In *Artificial Intelligence and Statistics*, pages 27–36. PMLR, 2012.

Noga Alon, Shay Moran, and Amir Yehudayoff. Sign rank versus VC dimension. In *Conference on Learning Theory*, pages 47–80, 2016.

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

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

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

Krishnakumar Balasubramanian. Nonparametric modeling of higher-order interactions via hypergraphons. *Journal of Machine Learning Research*, 22:1–25, 2021.

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

Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pages 1863–1874, 2019.

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

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Eric C Chi, Brian J Gaines, Will Wei Sun, Hua Zhou, and Jian Yang. Provable convex co-clustering of tensors. *Journal of Machine Learning Research*, 21(214):1–58, 2020.

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

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

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

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

Ravi Ganti, Nikhil Rao, Laura Balzano, Rebecca Willett, and Robert Nowak. On learning high dimensional structured single index models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 1898–1904, 2017.

Ravi Sastry Ganti, Laura Balzano, and Rebecca Willett. Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, pages 1873–1881, 2015.

Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

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

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

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

Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, 8:2265–2295, 2007.

Rungang Han, Rebecca Willett, and Anru R Zhang. *The Annals of Statistics*, 50(1):1–29, 2022.

Botao Hao, Boxiang Wang, Pengyuan Wang, Jingfei Zhang, Jian Yang, and Will Wei Sun. Sparse tensor additive regression. *Journal of machine learning research*, 22, 2021.

Lifang He, Chun-Ta Lu, Guixiang Ma, Shen Wang, Linlin Shen, S Yu Philip, and Ann B Ragin. Kernelized support tensor machines. In *International Conference on Machine Learning*, pages 1442–1451, 2017.

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

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

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

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

Jiaxin Hu, Chanwoo Lee, and Miaoyan Wang. Generalized tensor decomposition with features on multiple modes. *Journal of Computational and Graphical Statistics*, 31(1): 204–218, 2022.

Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.

Dmitri A Jdanov, Domantas Jasilionis, Vladimir M Shkolnikov, and Magali Barbieri. Human mortality database. *Encyclopedia of gerontology and population aging/editors Danan Gu, Matthew E. Dupre. Cham: Springer International Publishing, 2020*, 2019.

Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

Tamara G Kolda. A counterexample to the possibility of an extension of the eckart–young low-rank approximation theorem for the orthogonal rank tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 24(3):762–767, 2003.

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

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

Chanwoo Lee and Miaoyan Wang. Tensor denoising and completion based on ordinal observations. In *International Conference on Machine Learning*, pages 5778–5788, 2020.

Chanwoo Lee, Lexin Li, Hao Helen Zhang, and Miaoyan Wang. Nonparametric trace regression in high dimensions via sign series representation. *arXiv preprint arXiv:2105.01783*, 2021.

Yonghui Li, Yong Liu, Jun Li, Wen Qin, Kuncheng Li, Chunshui Yu, and Tianzi Jiang. Brain anatomical network and intelligence. *PLoS Comput Biol*, 5(5):e1000395, 2009.

László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

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

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

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

Greg Ongie, Rebecca Willett, Robert D. Nowak, and Laura Balzano. Algebraic variety models for high-rank matrix completion. In *International Conference on Machine Learning*, pages 2691–2700, 2017.

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

Clayton Scott. Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In *International Conference on Machine Learning*, 2011.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, 22:580–615, 1994.

Xiaotong Shen, George C Tseng, Xuegong Zhang, and Wing Hung Wong. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Lu Wang, Daniele Durante, Rex E Jung, and David B Dunson. Bayesian network–response regression. *Bioinformatics*, 33(12):1859–1866, 2017.

Lu Wang, Zhengwu Zhang, and David Dunson. Common and individual structure of brain networks. *The Annals of Applied Statistics*, 13(1):85–112, 2019.

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

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

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

Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *International Conference on Machine Learning*, pages 10544–10554, 2020.

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

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

Jingfei Zhang, Will Wei Sun, and Lexin Li. Generalized connectivity matrix response regression with applications in brain connectivity studies. *Journal of Computational and Graphical Statistics*, 2022.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783, 2017.

Zhengwu Zhang, Genevera I Allen, Hongtu Zhu, and David Dunson. Tensor network factorizations: Relationships between brain structural connectomes and traits. *Neuroimage*, 197:330–343, 2019.

Yufei Zhao. Hypergraph limits: a regularity approach. *Random Structures & Algorithms*, 47 (2):205–226, 2015.

Ya Zhou, Raymond KW Wong, and Kejun He. Broadcasted nonparametric tensor regression. *arXiv preprint arXiv:2008.12927*, 2020.

Appendix A. Experimental details in Section 1.1

In Section 1.1 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 experiment 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. We define two numerical ranks; CP and Tucker rank.

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

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

We compute $\hat{r}_{\text{cp}}(\Theta)$ by searching over $s \in \{1, \ldots, 30^2\}$, where for each s, we (approximately) solve the least-square minimization using built-in `cp` function in R package `rTensor` with default setting (iteration = 25, tolerance = 10^{-5}). We compute $\hat{r}_{\text{tucker}}(\Theta)$ by searching over $s \in \{1, \ldots, 30\}$ and solving the least-square minimization for each s using built-in `tucker` function in R package `rTensor` with default setting (iteration = 25, tolerance = 10^{-5}). We repeat steps 1-2 ten times, and plot the averaged numerical rank of Θ versus transformation level c in Figure 1a.

Appendix B. Proofs of theory in Section 3

B.1 Proofs of Propositions 1-3

Proof of Proposition 1. Part (a). By definition of sign-rank (4),

$$\text{srank}(\Theta - \pi) \le \text{srank}(\Theta) + \text{srank}(\mathbf{1} \otimes \cdots \otimes \mathbf{1}) = \text{srank}(\Theta) + 1 \le \text{rank}(\Theta) + 1.$$

Part (b). The strict 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 $x \mapsto g(x) - g(0)$ is sign preserving. Specifically, if $x \ge 0$, then $g(x) \ge g(0)$. Conversely, if $g(x) \ge g(0)$, then applying g^{-1} to both sides gives $x \ge 0$.

When g is strictly decreasing, the mapping $x \mapsto g(x) - g(0)$ is sign reversing. Specifically, if $x \ge 0$, then $g(x) \le g(0)$. Conversely, if $g(x) \le g(0)$, then applying g^{-1} to both sides gives $x \le 0$.

Combining the above two cases gives that $\Theta \simeq g(\Theta) - g(0)$ or $\Theta \simeq -(g(\Theta) - g(0))$. Since constant multiplication does not change the tensor rank, we have $\text{srank}(\Theta) = \text{srank}(g(\Theta) - g(0)) \le \text{rank}(g(\Theta)) + 1$.

Part (c). See Example 6 for constructive examples.

□

Proof of Proposition 2. We reorder the tensor indices along each mode such that $x_1^{(k)} \le \cdots \le x_{d_k}^{(k)}$ 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

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

  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 $\operatorname{sgn}(\Theta - \pi) = -1$,

$$\begin{aligned} \{\omega \colon \Theta(\omega) - \pi < 0\} &= \{\omega \colon g(\mathcal{Z}_{\max}(\omega)) < \pi\} \\ &= \cup_{s \in \mathcal{I}} \{\omega \colon \mathcal{Z}_{\max}(\omega) \in (z_s, z_{s+1})\} \\ &= \cup_{s \in \mathcal{I}} \Big(\{\omega \colon x^{(k)}_{i_k} < z_{s+1} \text{ for all } k \in [K]\} \cap \{\omega \colon x^{(k)}_{i_k} \leq z_s \text{ for all } k \in [K]\}^c \Big). \end{aligned} \tag{25}$$

  The equation (25) is equivalent to

$$\mathbb{1}(\Theta(i_1, \ldots, i_K) < \pi) = \sum_{s \in \mathcal{I}} \left(\prod_k \mathbb{1}(x^{(k)}_{i_k} < z_{s+1}) - \prod_k \mathbb{1}(x^{(k)}_{i_k} \leq z_s) \right), \tag{26}$$

  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 (26) implies the low-rank representation of $\operatorname{sgn}(\Theta - \pi)$,

$$\operatorname{sgn}(\Theta - \pi) = 1 - 2 \sum_{s \in \mathcal{I}} \left(\boldsymbol{a}^{(1)}_{s+1} \otimes \cdots \otimes \boldsymbol{a}^{(K)}_{s+1} - \bar{\boldsymbol{a}}^{(1)}_s \otimes \cdots \otimes \bar{\boldsymbol{a}}^{(K)}_s \right), \tag{27}$$

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

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

  Therefore, by (27) and the assumption $|\mathcal{I}| \leq r - 1$, we conclude that

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

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

Proof of Proposition 3. The conclusion directly follows from the definition of tensor rank. □

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 $i^* \in \{2, \dots, d\}$, 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

$$\mathrm{sgn}(\Theta(i,j) - \pi) = \begin{cases} -1, & \text{both } i \text{ and } j \text{ are smaller than } i^*; \\ 1, & \text{otherwise.} \end{cases} \tag{28}$$

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 $\mathrm{rank}(\Theta) \ge \mathrm{rank}(\Theta(:, :, 1, \dots, 1)) = d$. For the sign-rank with feasible π, notice that the sign tensor $\mathrm{sgn}(\Theta - \pi)$ takes the similar form as in (28),

$$\mathrm{sgn}(\Theta(i_1, \dots, i_K) - \pi) = \begin{cases} -1, & i_k < i^* \text{ for all } k \in [K]; \\ 1, & \text{otherwise,} \end{cases} \tag{29}$$

where i^* denotes the index that satisfies $\log(1+\frac{i^*-1}{d}) < \pi \le \log(1+\frac{i^*}{d})$. The equation (29) implies that $\mathrm{sgn}(\Theta - \pi) = -2\boldsymbol{a}^{\otimes K} + 1$, where $\boldsymbol{a} = (1, \dots, 1, 0, \dots, 0)^T$ takes 1 on the i-th entry if $i < i^*$ and 0 otherwise. Henceforth $\mathrm{srank}(\Theta - \pi) = 2$. □

Remark 13. Example 5 is proved in the similar way.

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 $\boldsymbol{b} = (2^{-1}, 2^{-2}, \dots, 2^{-d})^T \in \mathbb{R}^d$ and $\operatorname{rev}(\boldsymbol{b}) = (2^{-d}, \dots, 2^{-1})^T \in \mathbb{R}^d$. We construct the following matrix

$$\boldsymbol{A} = \boldsymbol{b} \otimes \operatorname{rev}(\boldsymbol{b}) + \operatorname{rev}(\boldsymbol{b}) \otimes \boldsymbol{b}. \tag{30}$$

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.,

$$\boldsymbol{A}(i,j) \ge \boldsymbol{A}(i',j'), \quad \text{for all } |i-j| \ge |i'-j'|. \tag{31}$$

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 (31). By definition (30), $\boldsymbol{A}$ is a rank-2 matrix. Henceforce, $\operatorname{srank}(\boldsymbol{M}-\pi) = \operatorname{srank}(\boldsymbol{A}-\pi') \le 3$. □

Appendix C. Proofs of Theory in Section 4

C.1 Proof of Theorem 1

Proof of Theorem 1. Fix $\pi \in [-1,1]$. Denote $\bar{\Theta} = (\Theta - \pi)$ and $\bar{\mathcal{Y}} = (\mathcal{Y} - \pi)$. 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

$$\operatorname{Risk}(\mathcal{Z}) - \operatorname{Risk}(\bar{\Theta}) = \mathbb{E}_{\omega\sim\Pi} \underbrace{\mathbb{E}_{\mathcal{Y}(\omega)}\left\{|\mathcal{Y}(\omega)-\pi|\left[\left|\mathcal{Z}(\omega) - \operatorname{sgn}(\bar{\mathcal{Y}}(\omega))\right| - \left|\bar{\Theta}(\omega) - \operatorname{sgn}(\bar{\mathcal{Y}}(\omega))\right|\right]\right\}}_{\stackrel{\text{def}}{=} I(\omega)}. \tag{32}$$

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

$$\begin{aligned} I(\omega) &= \mathbb{E}_{y|\omega}\left[(y-\pi)(\bar{\theta}-z)\mathbb{1}(y \ge \pi) + (\pi - y)(z-\bar{\theta})\mathbb{1}(y < \pi)\right] \\ &= \mathbb{E}_{y|\omega}\left[(\bar{\theta}-z)(y-\pi)\right] \\ &= \left[\operatorname{sgn}(\theta-\pi) - z\right](\theta-\pi) \\ &= |\operatorname{sgn}(\theta-\pi) - z||\theta-\pi| \ge 0, \end{aligned} \tag{33}$$

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 (33) is attained when $z = \operatorname{sgn}(\theta-\pi)$ or $\theta = \pi$. Combining (33) with (40), we conclude that, for all $\mathcal{Z} \in \{-1,1\}^{d_1\times\cdots\times d_K}$,

$$\operatorname{Risk}(\mathcal{Z}) - \operatorname{Risk}(\bar{\Theta}) = \mathbb{E}_{\omega\sim\Pi}|\operatorname{sgn}(\Theta(\omega)-\pi) - \mathcal{Z}(\omega)||\Theta(\omega)-\pi| \ge 0. \tag{34}$$

In particular, setting $\mathcal{Z} = \bar{\Theta} = \operatorname{sgn}(\Theta-\pi)$ in (34) yields the minimum. Therefore,

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

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

C.2 Proof of Theorem 2

Proof of Theorem 2. Fix $\pi \notin \mathcal{N}$ and denote $\bar{\Theta} = (\Theta - \pi)$. Based on (34) in Theorem 1, we have

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) = \mathbb{E}\left[|\text{sgn}\mathcal{Z} - \text{sgn}\bar{\Theta}||\bar{\Theta}|\right]. \tag{35}$$

The Definition 3 states that

$$\mathbb{P}\left(|\bar{\Theta}| \le t\right) \le \begin{cases} ct^{\alpha}, & \text{for all } \Delta s \le t < \rho(\pi, \mathcal{N}), \\ C\Delta s, & \text{for all } 0 \le t < \Delta s. \end{cases} \tag{36}$$

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 (35), for all $0 \le t < \rho(\pi, \mathcal{N})$,

$$\begin{aligned} \text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) &\ge t\mathbb{E}\left(|\text{sgn}\mathcal{Z} - \text{sgn}\bar{\Theta}|\mathbb{1}\{|\bar{\Theta}| > t\}\right) \\ &\ge 2t\mathbb{P}\left(\text{sgn}\mathcal{Z} \ne \text{sgn}\bar{\Theta} \text{ and } |\bar{\Theta}| > t\right) \\ &\ge 2t\Big\{\mathbb{P}\left(\text{sgn}\mathcal{Z} \ne \text{sgn}\bar{\Theta}\right) - \mathbb{P}\left(|\bar{\Theta}| \le t\right)\Big\} \\ &\ge t\Big\{\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) - C\Delta s - 2ct^{\alpha}\Big\}, \end{aligned} \tag{37}$$

  where the last line follows from the definition of MAE and (36). We maximize the lower bound (37) with respect to t, and obtain the optimal t_{opt},

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

  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 (37) becomes

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

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

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

  where $C(\pi) > 0$ is a multiplicative factor independent of $\mathcal{Z}$.
- Case 2: $\alpha = \infty$. The inequality (37) now becomes

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) \ge t\left[\text{MAE}(\text{sgn}\bar{\Theta}, \text{sgn}\mathcal{Z}) - C\Delta s\right], \quad \text{for all } 0 \le t < \rho(\pi, \mathcal{N}). \tag{38}$$

  The conclusion follows by taking $t = \frac{\rho(\pi, \mathcal{N})}{2}$ in the inequality (38).

□

Remark 14. The proof of Theorem 2 shows that, under global α-smoothness of Θ,

$$\mathrm{MAE}(\mathrm{sgn}\mathcal{Z}, \mathrm{sgn}\bar{\Theta}) \lesssim [\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})} \left[\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})\right] + \Delta s, \quad (39)$$

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.

C.3 Proof of Theorem 3

Proof of Theorem 3. Denote $\bar{\Theta} = (\Theta - \pi)$ and $\bar{\mathcal{Y}} = (\mathcal{Y} - \pi)$. Similar to the proof of Theorem 1, we evaluate the excess risk

$$\mathrm{Risk}_q(\mathcal{Z}) - \mathrm{Risk}_q(\bar{\Theta}) = \mathbb{E}_{\omega \sim \Pi} \underbrace{\mathbb{E}_{\mathcal{Y}(\omega)} \left\{ |\mathcal{Y}(\omega) - \pi|^q \left[\left| \mathcal{Z}(\omega) - \mathrm{sgn}(\bar{\mathcal{Y}}(\omega)) \right| - \left| \bar{\Theta}(\omega) - \mathrm{sgn}(\bar{\mathcal{Y}}(\omega)) \right| \right] \right\}}_{\stackrel{\mathrm{def}}{=} I(\omega)}. \quad (40)$$

The expression of $I(\omega)$ is simplified as

$$\begin{aligned} I(\omega) = \mathbb{E}_{\mathcal{Y}(\omega) \geq \pi} \left[(\mathcal{Y}(\omega) - \pi)^q (\mathrm{sgn}(\bar{\Theta}(\omega)) - \mathrm{sgn}(\mathcal{Z}(\omega))) \right] \\ + \mathbb{E}_{\mathcal{Y}(\omega) < \pi} \left[(\pi - \mathcal{Y}(\omega))^q (\mathrm{sgn}(\mathcal{Z}(\omega)) - \mathrm{sgn}(\bar{\Theta}(\omega))) \right]. \end{aligned}$$

We consider two cases when q is an odd number and an even number.

- Case 1: q is an odd number. In this case, $I(\omega)$ becomes

$$\begin{aligned} I(\omega) &= \mathbb{E}_{\mathcal{Y}(\omega)} \left[(\mathcal{Y}(\omega) - \pi)(\mathrm{sgn}(\bar{\Theta}(\omega)) - \mathrm{sgn}(\mathcal{Z}(\omega))) \right] \\ &= \left(\mathrm{sgn}(\bar{\Theta}(\omega)) - \mathrm{sgn}(\mathcal{Z}(\omega)) \right) \mathbb{E}_{\mathcal{Y}(\omega)} \left[\mathcal{Y}(\omega) - \pi \right]^q \end{aligned}$$

  Therefore, $I(\omega)$ is minimized when

$$\mathrm{sgn}(\mathcal{Z}(\omega)) = \mathrm{sgn}(\mathbb{E}_{\mathcal{Y}(\omega)} (\mathcal{Y}(\omega) - \pi)^q).$$

  Denote $y = \mathcal{Y}(\omega)$, $\theta = \Theta(\omega)$, and $\mathcal{E}(\omega) = \epsilon$. Under the condition that ϵ is symmetric zero-mean noise, we show that

$$\mathrm{sgn}\left(\mathbb{E}_y (y - \pi)^q\right) = \mathrm{sgn}(\theta - \pi). \quad (41)$$

  Notice that

$$\begin{aligned} \mathbb{E}_y (y - \pi)^q &= \sum_{k=0}^{q} \binom{q}{k} (\theta - \pi)^{p-k} \mathbb{E}_y (y - \theta)^k \\ &= \sum_{k=0}^{(p-1)/2} (\theta - \pi)^{p-2k} \mathbb{E} \epsilon^{2k}, \end{aligned}$$

  where the last line uses the noise distribution is symmetric about the origin. Therefore, we prove (41).

- Case 2: q is an even number. In this case, $I(\omega)$ becomes

$$I(\omega) = \left(\text{sgn}(\bar{\Theta}(\omega)) - \text{sgn}(\mathcal{Z}(\omega))\right)\left(\mathbb{E}_{\mathcal{Y}(\omega)\geq\pi}\left[\mathcal{Y}(\omega)-\pi\right]^q - \mathbb{E}_{\mathcal{Y}(\omega)<\pi}\left[\mathcal{Y}(\omega)-\pi\right]^q\right).$$

Therefore, $I(\omega)$ is minimized when

$$\text{sgn}(\mathcal{Z}(\omega)) = \text{sgn}\left(\mathbb{E}_{\mathcal{Y}(\omega)\geq\pi}\left[\mathcal{Y}(\omega)-\pi\right]^q - \mathbb{E}_{\mathcal{Y}(\omega)<\pi}\left[\mathcal{Y}(\omega)-\pi\right]^q\right). \tag{42}$$

Denote $y = \mathcal{Y}(\omega)$, $\theta = \Theta(\omega)$, and $\mathcal{E}(\omega) = \epsilon$. Under the condition that ϵ is symmetric zero-mean noise, we show that

$$\text{sgn}\left(\mathbb{E}_{y-\pi\geq 0}\,(y-\pi)^q - \mathbb{E}_{y-\pi<0}\,(y-\pi)^q\right) = \text{sgn}(\theta-\pi).$$

Let $x := y-\pi = \theta-\pi+\epsilon$. Notice that x is a random variable which is symmetric at $m := \theta-\pi$. First suppose $m \geq 0$, then we have

$$\begin{aligned}
\mathbb{E}_{y-\pi\geq 0}\,(y-\pi)^q - \mathbb{E}_{y-\pi<0}\,(y-\pi)^q &= \int_{x\geq 0} x^q dx - \int_{x<0} x^q dx \\
&\geq \int_{x\geq m} x^q dx - \int_{x<m} x^q dx \\
&= \int_{x\geq m} \sum_{k=0}^{q} \binom{q}{k} (x-m)^k m^{p-k} dx \\
&\qquad - \int_{x<m} \sum_{k=0}^{q} \binom{q}{k} (x-m)^k m^{p-k} dx \\
&= 2\int_{x\geq m} \sum_{k=0}^{q/2-1} \binom{q}{2k+1} (x-m)^{2k+1} m^{p-2k-1} \\
&\geq 0,
\end{aligned}$$

where the last equality uses the symmetry of x around m. We can prove the case $m<0$ in the same way. Therefore, we show that (42) holds true.

□

Appendix D. Proof of Theory in Section 5

D.1 Proof of sign tensor estimation

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 6.

Lemma 1 (Variance-to-mean relationship)**.** Consider the same setup as in Theorem 4. Fix $\pi \notin \mathcal{N}$. Let $L(\mathcal{Z}, \bar{Y}_\Omega)$ be the l_1 weighted classification loss

$$L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) = \frac{1}{|\Omega|}\sum_{\omega\in\Omega} \underbrace{|\bar{\mathcal{Y}}(\omega)|}_{l_1 \text{ weight}} \times \underbrace{|\text{sgn}\mathcal{Z}(\omega) - \text{sgn}\bar{\mathcal{Y}}(\omega)|}_{\text{classification loss}}$$

$$= \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}), \tag{43}$$

where we have denoted the function $\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}) \stackrel{\text{def}}{=} |\bar{\mathcal{Y}}(\omega)||\text{sgn}\mathcal{Z}(\omega) - \text{sgn}\bar{\mathcal{Y}}(\omega)|$. Under Definition 3 of the α-smoothness of Θ, we have

$$\text{Var}[\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)] \lesssim [\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})}[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})] + \Delta s, \tag{44}$$

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 $\mathcal{Y}(\omega)$ and $\omega \sim \Pi$.

Proof of Lemma 1. We expand the variance by

$$\begin{aligned} \text{Var}[\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)] &\lesssim \mathbb{E}|\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)|^2 \\ &\lesssim \mathbb{E}|\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)| \\ &\le \mathbb{E}|\text{sgn}\mathcal{Z} - \text{sgn}\bar{\Theta}| = \text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}), \end{aligned} \tag{45}$$

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 (44) then directly follows by applying Remark 14 to (45). □

Proof of bound (17) *in Theorem 4.* 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

$$\hat{\mathcal{Z}} = \mathop{\arg\min}_{\mathcal{Z}:\, \text{rank}(\mathcal{Z}) \le r, \|\mathcal{Z}\|_F \le 1} L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega). \tag{46}$$

We next use the local empirical process theory to bound $\hat{\mathcal{Z}}$. To facilitate the analysis, we view the data $\bar{\mathcal{Y}}_\Omega = \{\bar{\mathcal{Y}}(\omega) : \omega \in \Omega\}$ 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 (43) becomes

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

We use $f_{\mathcal{Z}} : [d_1] \times \cdots \times [d_n] \to \mathbb{R}$ 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

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

is an empirical process induced by function $f_{\mathcal{Z}} \in \mathcal{F}_{\mathcal{T}}$ where $\mathcal{T} = \{\mathcal{Z} : \text{rank}(\mathcal{Z}) \le r,\ \|\mathcal{Z}\|_F \le 1\}$. 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 (46), we have,

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

Therefore, we have the following inclusion of probability events,

$$\begin{aligned}
&\left\{(\omega, \mathcal{Y}_\omega)\colon \mathrm{Risk}(\hat{\mathcal{Z}}) - \mathrm{Risk}(\bar{\Theta}) \geq L_n\right\} \\
&\subset \left\{(\omega, \mathcal{Y}_\omega)\colon \exists \mathcal{Z} \text{ s.t. } \mathrm{rank}(\mathcal{Z}) \leq r, \mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta}) \geq L_n, \text{ and } \frac{1}{n}\sum_{i=1}^{n} \Delta_i(f_{\mathcal{Z}}, \bar{\Theta}) \leq 0\right\} \\
&\subset \left\{(\omega, \mathcal{Y}_\omega)\colon \sup_{\substack{\mathrm{rank}(\mathcal{Z}) \leq r \\ \mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta}) \geq L_n}} -\frac{1}{n}\sum_{i=1}^{n} \Delta_i(f_{\mathcal{Z}}, \bar{\Theta}) \geq 0\right\} \\
&\subset \bigcup_{\ell=1}^{\infty} \left\{(\omega, \mathcal{Y}_\omega)\colon \sup_{\mathcal{Z} \in A_\ell} -\frac{1}{n}\sum_{i=1}^{n} \Delta_i(f_{\mathcal{Z}}, \bar{\Theta}) \geq 0\right\}, \qquad (47)
\end{aligned}$$

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

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

for $\ell = 1, 2, \ldots$. Let Γ denote the target probability for the first line in (47). To bound Γ, we bound the sum of probability over the sets A_ℓ. For each A_ℓ, we consider the centered empirical process,

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

Notice $(\ell+1)L_n \geq \mathbb{E}\Delta_i(f_{\mathcal{Z}}, \bar{\Theta}) = \mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta}) \geq \ell L_n$ for all $\mathcal{Z} \in A_\ell$. Combining (47), (48) and union bound yields

$$\Gamma \leq \sum_{\ell=1}^{\infty} \mathbb{P}\left\{\sup_{\mathcal{Z} \in A_\ell} v_n(f_{\mathcal{Z}}) \geq \ell L_n =: M(\ell)\right\}. \qquad (49)$$

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

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

We next bound the right-hand side of (49) by choosing L_n that satisfies conditions in Theorem 6 (The specification of L_n is deferred to the next paragraph). One such L_n is chosen, Theorem 6 gives us

$$\Gamma \lesssim \sum_{\ell=1}^{\infty} \exp\left(-\frac{nM^2(\ell)}{V(\ell) + 2M(\ell)}\right) \qquad (50)$$

$$\lesssim \sum_{\ell=1}^{\infty} \exp(-\rho \ell n L_n)$$
$$\leq \left(\frac{e^{-n\rho L_n}}{1 - e^{-n\rho L_n}} \right).$$

Now, we specify L_n that satisfies the condition of Theorem 6. The quantity L_n is determined by the solution to the following inequality,

$$\sup_{\ell \geq 1} \frac{1}{x} \int_{x}^{\sqrt{x^{\alpha/(\alpha+1)} + x/\rho + \Delta s}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}_{\mathcal{T}}, \|\cdot\|_2)} d\varepsilon \lesssim n^{1/2}, \quad \text{where } x = \ell L_n. \tag{51}$$

In particular, the smallest L_n satisfying (51) 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 4) for function family $\mathcal{F}_{\mathcal{T}}$.

Based on Lemma 2, the inequality (51) is satisfied with the choice

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

Finally, it follows from Theorem 6 and (50) that

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

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 (39) gives that, with a high probability at least $1 - \exp(-n t_n)$,

$$\begin{aligned} \text{MAE}(\text{sgn}\hat{\mathcal{Z}}, \text{sgn}\bar{\Theta}) &\lesssim [\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta})]^{\alpha/(\alpha+1)} + \frac{1}{\rho}[\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta})] + \Delta s \\ &\lesssim t_n^{\alpha/(\alpha+2)} + \frac{1}{\rho^{\alpha/\alpha+1}} t_n^{\alpha/(\alpha+1)} + \frac{1}{\rho} t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho^2} t_n \\ &\leq 4 t_n^{\alpha/(\alpha+2)} + \frac{4}{\rho^2} t_n, \end{aligned} \tag{52}$$

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 (52) and absorb the term K into the constant. The conclusion is then proved by noting $n = |\Omega|$ by definition. □

Definition 4 (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

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

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 $\mathcal{T} = \{\mathcal{Z} \in \mathbb{R}^{d_1\times\cdots\times d_K} : \text{rank}(\mathcal{Z}) \le r,\ \|\mathcal{Z}\|_F \le 1\}$ 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 $\Delta s \lesssim n^{-1}$,

$$\sup_{\ell\ge 1} \frac{1}{\ell L_n} \int_{\ell L_n}^{\sqrt{\ell L_n^{\alpha/(\alpha+1)} + \frac{\ell L_n}{\rho(\pi,\mathcal{N})} + \Delta s}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}_\mathcal{T}, \|\cdot\|_2)} d\varepsilon \le C n^{1/2}, \tag{53}$$

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 (53) gives

$$g(L, \ell) = \frac{1}{\ell L} \int_{\ell L}^{\sqrt{\ell L^{\alpha/(\alpha+1)} + \rho^{-1}\ell L + \Delta s}} \sqrt{d_{\max} r K \log\left(\frac{K}{\varepsilon}\right)} d\varepsilon. \tag{54}$$

Define $g(L) := \sup_{\ell\ge 1} g(L, \ell)$. By the monotonicity the integrand in (54), we bound $g(L)$ by

$$\begin{aligned} g(L) &\le \sup_{\ell\ge 1} \frac{\sqrt{d_{\max} r K}}{\ell L} \int_{\ell L}^{\sqrt{\ell L^{\alpha/(\alpha+1)} + \rho^{-1}\ell L + n^{-1}}} \sqrt{\log\left(\frac{K}{\ell L}\right)} d\varepsilon \\ &\le \sup_{\ell\ge 1} \sqrt{d_{\max} r K \log\left(\frac{K}{\ell L}\right)} \left(\frac{(\ell L)^{\alpha/(2\alpha+2)} + \sqrt{\rho^{-1}\ell L + n^{-1}}}{\ell L} - 1\right) \\ &\lesssim \sqrt{d_{\max} r K \log(1/L)} \left[\frac{1}{L^{(\alpha+2)/(2\alpha+2)}} + \frac{1}{\sqrt{\rho L}}\left(1 + \frac{\rho}{2nL}\right)\right], \end{aligned} \tag{55}$$

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 (53). Plugging L_n into the last line of (55) 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 6 (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

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

D.2 Proof of signal estimation error

Proof of bound (18) *in Theorem 4.* By definition of $\hat{\Theta}$, we have

$$\begin{aligned} \mathrm{MAE}(\hat{\Theta}, \Theta) &= \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn} \hat{Z}_\pi - \Theta \right| \\ &\le \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \Big(\mathrm{sgn} \hat{Z}_\pi - \mathrm{sgn}(\Theta - \pi) \Big) \right| + \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}(\Theta - \pi) - \Theta \right| \\ &\le \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{MAE}(\mathrm{sgn} \hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)) + \frac{1}{H}, \end{aligned} \tag{56}$$

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 (56). For any given $t \geq t_n = \frac{d_{\max} r K \log n}{n}$, 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,

$$\frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)) \lesssim t^{\alpha/(\alpha+2)} + \frac{1 + |\mathcal{N}|_{\text{jump}}}{H} + Ht. \tag{57}$$

The bound (17) 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 $\mathcal{H}_1 \stackrel{\text{def}}{=} \{\pi \in \mathcal{H} \colon \rho(\pi, \mathcal{N}) < \frac{1}{H}\}$ and $\mathcal{H}_2 = \mathcal{H} \setminus \mathcal{H}_1$. Notice $|\mathcal{H}_1|_{\text{jump}} \leq 2|\mathcal{N}|_{\text{jump}}$. We expand the left hand side of (57) by

$$\begin{aligned} &\frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)) \\ =&\frac{1}{2H+1} \sum_{\pi \in \mathcal{H}_1} \mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)) + \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}_2} \mathrm{MAE}(\mathrm{sgn}\hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)). \end{aligned} \tag{58}$$

The first term involves only $2|\mathcal{N}|_{\text{jump}}$ many number of sumnmands thus can be bounded by $4|\mathcal{N}|_{\text{jump}}/(2H+1)$. 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

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

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 (58) completes the proof for conclusion (57); that is

$$\mathbb{P}\left(\mathrm{MAE}(\hat{\Theta}, \Theta) \lesssim t^{\alpha/(\alpha+2)} + \frac{1 + |\mathcal{N}|_{\text{jump}}}{H} + Ht \right) \geq \mathbb{P}(A). \tag{59}$$

Based on the proof of sign error bound (17) and union bound over $\pi \in \mathcal{H}$, we have, for all $t \geq t_n$,

$$\mathbb{P}(A) \geq 1 - \sum_{\pi \in \mathcal{H}} \mathbb{P}\left(\mathrm{MAE}(\mathrm{sgn}\hat{\mathcal{Z}}_\pi, \mathrm{sgn}(\Theta - \pi)) \gtrsim t^{\alpha/(\alpha+2)} + \frac{t}{\rho(\pi, \mathcal{N})^2} \right)$$

$$\gtrsim 1 - (2H+1)\exp(-nt) \gtrsim 1 - \exp(-nt + \log H). \tag{60}$$

We choose $t \asymp t_n \log H$ in (60) so that $\log H$ is negligible compared to nt. Finally, combining (59) and (60) with the choice of t yields

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

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

□

Lemma 3. Fix $\pi' \in \mathcal{N}$ and a sequence $\Pi = \{-1, \ldots, -1/H, 0, 1/H, \ldots, 1\}$ 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 $x \geq 1$. □

D.3 Proof of Corollary 1

The conclusion readily follows from bound (18) in Theorem 4.

Appendix E. Proofs of Theory in Section 6

E.1 Proof of Corollary 2

Proof of Corollary 2. First we prove the sign tensor estimation error, i.e., for all $\pi \notin \mathcal{N}$,

$$\mathrm{MAE}(\mathrm{sgn}(\hat{\mathcal{Z}}_\pi), \mathrm{sgn}(\Theta - \pi)) \lesssim t_d^{\frac{\alpha}{\alpha+2}} + \frac{t_d}{\rho^2(\pi, \mathcal{N})}, \text{ where } t_d := \frac{r\sigma^2 d_{\max} \log d_{\max} \log|\Omega|}{|\Omega|} \tag{61}$$

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

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

We divide the sample space into two exclusive events:

- Event I: $\|\mathcal{E}\|_\infty \geq \sqrt{4\sigma^2 K \log d_{\max}}$;
- 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 4. We summarize the key difference compared to Section C. We expand the variance by

$$\begin{aligned}
\mathrm{Var}\left[\ell_\omega\left(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega\right) - \ell_\omega\left(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega\right)\right] &\leq \mathbb{E}|\ell_\omega(\mathcal{Z}(\omega), \bar{\mathcal{Y}}(\omega)) - \ell_\omega(\bar{\Theta}(\omega), \bar{\mathcal{Y}}(\omega))|^2 \\
&= \mathbb{E}|\bar{\mathcal{Y}}(\omega) - \bar{\Theta}(\omega) + \bar{\Theta}(\omega)|^2 |\mathrm{sgn}\mathcal{Z}(\omega) - \mathrm{sgn}\bar{\Theta}(\omega)| \\
&\leq 2\left(4\sigma^2 K \log d_{\max} + 2\right)\mathbb{E}|\mathrm{sgn}\mathcal{Z} - \mathrm{sgn}\bar{\Theta}| \\
&\lesssim (\sigma^2 K \log d_{\max})\mathrm{MAE}(\mathrm{sgn}\mathcal{Z}, \mathrm{sgn}\bar{\Theta}), \qquad (62)
\end{aligned}$$

where the third line uses the facts $\|\bar{\Theta}\|_\infty \leq 2$ 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 (62), the α-smoothness of Θ implies that for all measurable functions $f_{\mathcal{Z}}$, we have

$$\mathrm{Var}\Delta_i(f_{\mathcal{Z}}, \bar{\Theta}) \lesssim \left(\sigma^2 K \log d_{\max}\right)\left\{\left[\mathbb{E}\Delta_i(f_{\mathcal{Z}}, \bar{\Theta})\right]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho}\mathbb{E}\Delta_i(f_{\mathcal{Z}}, \bar{\Theta}) + \Delta s\right\}. \qquad (63)$$

Based on the proof of Theorem 4, the empirical process with variance-to-mean relationship (63) gives that

$$\mathbb{P}\left(\mathrm{Risk}(\hat{\mathcal{Z}}) - \mathrm{Risk}(\bar{\Theta}) \geq L_n\right) \lesssim \exp(-nt_n), \qquad (64)$$

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

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho}t_n, \quad \text{with } t_n = \frac{r\sigma^2 d_{\max}\log d_{\max}\log n}{n}, \qquad (65)$$

where constants (possibly depending on K) have been absorbed into the $\asymp$ relationship. Combining (64) and (65), we obtain that, with a high probability,

$$\mathrm{Risk}(\hat{\mathcal{Z}}) - \mathrm{Risk}(\bar{\Theta}) \lesssim \left(\frac{r\sigma^2 d_{\max}\log d_{\max}\log|\Omega|}{|\Omega|}\right)^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi, \mathcal{N})}\left(\frac{r\sigma^2 d_{\max}\log d_{\max}\log|\Omega|}{|\Omega|}\right), \qquad (66)$$

Therefore, combining (66) and (52) completes the proof of (61). The signal tensor estimation error follows readily from the proof of Theorem 4 in Section D.2. □

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 $s > 0$,

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

Proof of Lemma 4. The conclusion follows from

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

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

E.2 Proofs of Corollary 3 and Corollary 4

The conclusion readily follows from Theorem 4.

E.3 Proof of Theorem 5

Proof of Theorem 5. Write $\bar{\mathcal{Y}} = \mathcal{Y} - \pi$, $\bar{\Theta} = \Theta - \pi$, and $n = |\Omega|$. Here we consider the estimation

$$\hat{\mathcal{Z}}_{\pi,F} = \operatorname*{arg\,min}_{\mathrm{rank}(\mathcal{Z})\le r} \sum_{\omega\in\Omega} |\bar{\mathcal{Y}}(\omega)| \times F(\mathcal{Z}(\omega)\mathrm{sgn}(\bar{\mathcal{Y}}(\omega)) + \lambda_\pi \|\mathcal{Z}\|_F^2, \tag{67}$$

where $\lambda_\pi > 0$ is the penalty parameter and F is the hinge loss satisfying Assumption 2.

We follow the same line of proof as in Theorem 4. We first prove the sign tensor estimation error. That is, for all $\pi \notin \mathcal{N}$,

$$\mathrm{MAE}(\mathrm{sgn}(\hat{\mathcal{Z}}_\pi),\ \mathrm{sgn}(\Theta - \pi)) \lesssim t_n^{\frac{\alpha}{2+\alpha}} + \frac{1}{\rho^2(\pi, \mathcal{N})} t_n. \tag{68}$$

The tensor estimation error (24) directly follows from sign tensor estimation error (68) and the proof of Theorem 4. Therefore, it suffices to prove (68). Our proof uses the same techniques used in the proof of Theorem 4. We summarize only the key difference.

Fix $\pi \notin \mathcal{N}$. For notational simplicity, we suppress the subscript π, and write $\hat{\mathcal{Z}}$, λ in place of $\hat{\mathcal{Z}}_{\pi,F}$ and λ_π. Denote $n = |\Omega|$ and $\rho = \rho(\pi, \mathcal{N})$. Define $\ell_{\omega,F}(\mathcal{Z}) = |\bar{\mathcal{Y}}(\omega)| \times F(\mathcal{Z}(\omega)\mathrm{sgn}(\bar{\mathcal{Y}}(\omega))$ and $\ell_{\omega,F'}(\mathcal{Z}) = |\bar{\mathcal{Y}}(\omega)| \times F'(\mathcal{Z}(\omega)\mathrm{sgn}(\bar{\mathcal{Y}}(\omega))$ 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 $\mathcal{T} = \{\mathcal{Z} \colon \mathrm{rank}(\mathcal{Z}) \le r\}$,

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

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 (67), we have

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

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,

$$\begin{aligned}
&\left\{(\omega,\mathcal{Y}_\omega)\colon \mathrm{Risk}_{F'}(\hat{\mathcal{Z}}) - \mathrm{Risk}_{F'}(\bar{\Theta}) \ge 2L_n\right\}\\
\subset &\left\{(\omega,\mathcal{Y}_\omega)\colon \exists\mathcal{Z} \text{ s.t. } \mathrm{rank}(\mathcal{Z})\le r, \mathrm{Risk}_{F'}(\mathcal{Z}) - \mathrm{Risk}_{F'}(\bar{\Theta}) \ge 2L_n,\right.\\
&\qquad\qquad \left.\text{and } -\frac{1}{n}\sum_{i=1}^{n}\Delta_{i,F}(f_{\mathcal{Z}},\mathcal{Z}^{(n)}) + \lambda J^2 - \lambda\|\hat{\mathcal{Z}}\|_F^2 \ge 0\right\}\\
\overset{(*)}{\subset} &\left\{(\omega,\mathcal{Y}_\omega)\colon \exists\mathcal{Z} \text{ s.t. } \mathrm{rank}(\mathcal{Z})\le r, \mathrm{Risk}_{F'}(\mathcal{Z}) - \mathrm{Risk}_{F'}(\bar{\Theta}) \ge 2L_n,\right.\\
&\qquad\qquad \left.\text{and } -\frac{1}{n}\sum_{i=1}^{n}\Delta_{i,F'}(f_{\mathcal{Z}},\mathcal{Z}^{(n)}) + \lambda J^2 - \lambda\|\hat{\mathcal{Z}}\|_F^2 \ge 0\right\}\\
\subset &\left\{(\omega,\mathcal{Y}_\omega)\colon \sup_{\substack{\mathrm{rank}(\mathcal{Z})\le r\\ \mathrm{Risk}_{F'}(\mathcal{Z})-\mathrm{Risk}_{F'}(\bar{\Theta})\ge 2L_n}} -\frac{1}{n}\sum_{i=1}^{n}\Delta_{i,F'}(f_{\mathcal{Z}},\mathcal{Z}^{(n)}) + \lambda J^2 - \lambda\|\hat{\mathcal{Z}}\|_F^2 \ge 0\right\}\\
\subset &\bigcup_{\ell_1,\ell_2=1}^{\infty}\left\{(\omega,\mathcal{Y}_\omega)\colon \sup_{\mathcal{Z}\in A_{\ell_1,\ell_2}} -\frac{1}{n}\sum_{i=1}^{n}\Delta_{i,F'}(f_{\mathcal{Z}},\mathcal{Z}^{(n)}) + \lambda J^2 - \lambda\|\hat{\mathcal{Z}}\|_F^2 \ge 0\right\}, \qquad (69)
\end{aligned}$$

where $(*)$ comes from the fact

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

because the truncation constant $T = \max(2, J^2) \ge \max(2, \sup_n\|\mathcal{Z}^{(n)}\|_F^2)$. In the last line of (69), we have partitioned $\{\mathcal{Z}\colon \mathrm{rank}(\mathcal{Z})\le r \text{ and } \mathrm{Risk}_{F'}(\mathcal{Z}) - \mathrm{Risk}_{F'}(\bar{\Theta}) \ge 2L_n\}$ into union of A_{ℓ_1,ℓ_2} with

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

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

Let Γ denote the target probability for the first line in (69). For each A_{ℓ_1,ℓ_2}, we consider the centered empirical process,

$$v_n(f_{\mathcal{Z}}) := -\frac{1}{n}\sum_{i=1}^{n}\left(\Delta_{i,F'}(f_{\mathcal{Z}},\mathcal{Z}^{(n)}) - \mathbb{E}\Delta_{i,F'}(f_{\mathcal{Z}},\mathcal{Z}^{(n)})\right). \qquad (70)$$

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 (69), (70) and the union bound yields

$$\Gamma \leq \sum_{\ell_1,\ell_2=1}^{\infty} \mathbb{P}\left\{ \sup_{\mathcal{Z}\in A_{\ell_1,\ell_2}} v_n(f_{\mathcal{Z}}) \geq \ell_1 L_n + \lambda(\ell_2 - 2)J^2 =: M(\ell_1,\ell_2) \right\}. \tag{71}$$

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 6, we choose the pair (L_n, λ) satisfying

$$\sup_{\ell_1,\ell_2\geq 1} \frac{1}{x} \int_x^{\sqrt{x^{\alpha/(\alpha+1)} + x/\rho + \Delta s}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}_{\mathcal{T}}(\ell_2), \|\cdot\|_2)} d\varepsilon \lesssim n^{1/2}, \tag{72}$$

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 (72) as

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + \frac{t_n}{\rho}, \quad \text{and} \quad \lambda = \frac{L_n}{2J^2}, \tag{73}$$

where $t_n = \frac{d_{\max} r K \log n}{n}$. With the choice (73), we bound the right-hand side of (71) based on Theorem 6,

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

where the last line uses the fact that $2\rho\lambda J^2 = \rho L_n \gtrsim t_n \gtrsim n^{-1}$ from (73). The proof is then completed by a similar calculation as in (52). □