Beyond the Signs: Nonparametric Tensor Completion via Sign Series

Anonymous Authors[1]

Abstract

We consider the problem of tensor estimation from noisy observations with missing entries. A nonparametric approach to tensor completion is developed based on a new model which we coin as "sign representable tensors." The model represents a continuous-valued signal tensor using a series of sign tensors with low sign-ranks. Unlike earlier methods, the sign series representation effectively addresses both low- and high-rank signal tensors, while encompassing many existing tensor models—including CP models, Tucker models, single index models, certain hypergraphon models—as special cases. We show that the sign tensor series are theoretically characterized, and computationally solvable, by classification tasks with carefully-specified weights. The excess risk rate, estimation error bound, and sample complexity are established. The results uncover the joint contribution of statistical bias-variance errors and discretization errors. Numerical results demonstrate the robustness of our proposal over previous tensor methods.

1. Introduction

Advantages: (1) exactly recovers the signal tensor under a wide range of low- and high-rank tensor models; (2) brings the nonparametric advantages of flexibility into tensor estimation and completion; (3) achieves computational efficiency by leveraging classification and divide-and-conquer algorithms.

2. Preliminaries

We use the shorthand $[n]$ to denote the n-set $\{1, \dots, n\}$ for $n \in \mathbb{N}_+$. We use $\otimes$ to denote the outer product of vectors, $\|\boldsymbol{x}\|_2$ to denote the vector 2-norm, and $\mathbf{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d \colon \|\boldsymbol{x}\|_2 = 1\}$ to denote the $(d-1)$-dimensional unit sphere. Let $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K $(d_1, \dots, d_k)$-dimensional tensor, and $\mathcal{Y}(\omega) \in \mathbb{R}$ denote the tensor entry indexed by $\omega \in [d_1] \times \cdots \times [d_K]$. The Frobenius norm of $\mathcal{Y}$ is defined as $\|\mathcal{Y}\|_F = \sqrt{\sum_\omega \mathcal{Y}^2(\omega)}$. Unlike matrices, various notions of decomposition have been developed for tensors of order $K \geq 3$. The Canonical Polyadic (CP) tensor decomposition (Hitchcock, 1927) for a tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is defined as

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

where $\lambda_1 \geq \cdots \geq \lambda_r > 0$ are called tensor singular values, and $\boldsymbol{a}_s^{(k)} \in \mathbf{S}^{d_k - 1}$ are called tensor singular vectors, for all $s \in [r]$, $k \in [K]$. The minimal r for which the decomposition (1) holds is called the tensor rank, denoted as $\mathrm{rank}(\Theta)$.

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 $\mathrm{sgn}(\cdot)$ or general $f \colon \mathbb{R} \to \mathbb{R}$, to be applied to tensors in an element-wise manner. For a tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, its sign pattern $\mathrm{sgn}(\Theta)$ is an order-K $(d_1, \dots, d_K)$-dimensional binary tensor with entries in $\{-1, 1\}$.

3. Motivation and method overview

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

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

where $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the unknown signal tensor of interest, and $\mathcal{E}$ is a noise tensor consisting of mean-zero, independent but not necessarily identically distributed entries. We allow heterogenous noise in that the marginal distribution of noise entry $\mathcal{E}(\omega)$ may depend on ω. This incorporates, for example, a Bernoulli tensor whose entries $\mathcal{Y}(\omega)$ have mean $\Theta(\omega)$ and variance $\mathrm{var}(\mathcal{E}(\omega)) = \Theta(\omega)(1 - \Theta(\omega))$. In general, we assume the range of $\mathcal{Y}(\omega)$ is a bounded interval $[-A, A]$, $A > 0$; other than that, we make no particular parametric assumptions on the distribution of $\mathcal{E}(\omega)$.

Our observation is an incomplete data tensor from (2), denoted $\mathcal{Y}_\Omega$, where $\Omega \subset [d_1] \times \cdots \times [d_K]$ is the index set of

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

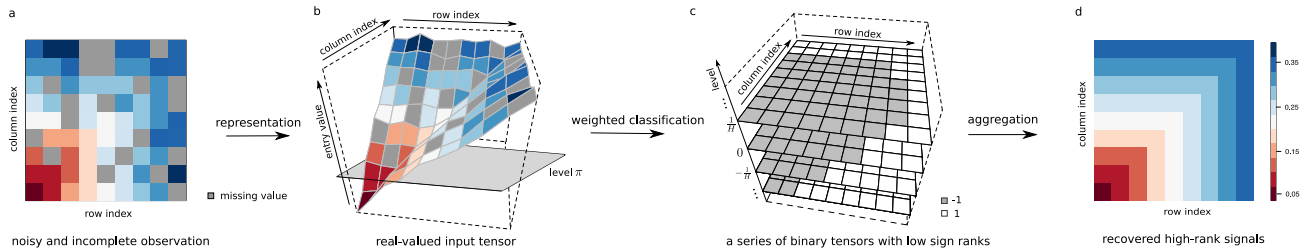

Figure 1. Illustration of our proposed method. For visualization purpose, we plot an order-2 tensor (a.k.a. matrix) in the figure; similar procedure applies to higher-order tensors. (a): input tensor $\mathcal{Y}_\Omega$ with noisy and incomplete entries. (b) and (c): our method uses weighted classification to estimate sign tensors $\mathrm{sgn}(\Theta - \pi)$ for a sequence of levels $\pi \in \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$. (d) output tensor $\hat{\Theta}$ with denoised and imputed entries. The depicted example is based on Example 5, where the true signal matrix has full rank.

observed entries. We consider a general model on Ω that allows both uniform and non-uniform samplings. Specifically, let $\Pi = \{\pi_\omega\}$ be an arbitrarily predefined probability distribution over the full index set with $\sum_{\omega \in [d_1 \times \cdots \times d_K]} \pi_\omega = 1$. We assume the entries ω in Ω are i.i.d. draws with replacement from the full index set using distribution Π. The sampling rule will be denoted as $\omega \sim \Pi$.

Our goal is to accurately estimate Θ from the incomplete, noisy observation $\mathcal{Y}_\Omega$. In particular, we focus on the following two problems:

- Q1 [Nonparametric tensor estimation]. How to flexibly estimate Θ under a wide range of structures, including both low-rankness and high-rankness?
- Q2 [Tensor completion]. How many observed tensor entries do we need to consistently estimate the signal Θ?

3.1. Inadequacies of common low-rank models

The signal plus noise model (2) is common in tensor literature. Most existing methods perform estimation based on the low-rankness of Θ (Anandkumar et al., 2014; Montanari and Sun, 2018; Kadmon and Ganguli, 2018; Cai et al., 2019). While these methods have shown great success in low-rank recovery, little is explored when the underlying signal tensor is of high rank. Here we provide two examples to illustrate the limitation of classical low-rank models.

The first example reveals the sensitivity of tensor rank to order-preserving transformations. Let $\mathcal{Z} \in \mathbb{R}^{30\times30\times30}$ be an order-3 tensor with $\mathrm{rank}(\mathcal{Z}) = 3$. Suppose a monotonic transformation $f(z) = (1 + \exp(-cz))^{-1}$ is applied to $\mathcal{Z}$ entrywise, and we observe data from model (2) with the signal tensor $\Theta = f(\mathcal{Z})$. Figure 2(a) plots the numerical rank of Θ versus c. Note that a smaller c implies an approximate linear transformation $f(z) \approx -cz$, whereas a larger c implies a higher nonlinearity $z \mapsto \{0, 1\}$. As we see, the rank increases rapidly with c, rending traditional low-rank tensor methods ineffective even in the presence of mild order-preserving nonlinearities. In applications of digital processing and genomics analysis, the tensor of interest often undergoes some unknown transformation prior to measurements. The sensitivity to transformation therefore makes the low-rank model less desirable in practice.

The second example shows that the classical low-rankness may exclude important specially-structured tensors. Here we consider the signal tensor of the form $\Theta = \log(1 + \mathcal{Z})$, where $\mathcal{Z}$ is an order-3 tensor with entries $\mathcal{Z}(i, j, k) = \frac{1}{d}\max(i, j, k)$ for $(i, j, k) \in [d]^3$. In this case neither Θ nor $\mathcal{Z}$ is low-rank; indeed, both tensors have rank lower bounded by dimension d as illustrated in Figure 2(b) (proofs in Appendix). The matrix analogy of this Θ was studied in Chan and Airoldi (2014) in the context of graphon analysis. However, classical low-rank models fail to address these types of structures.

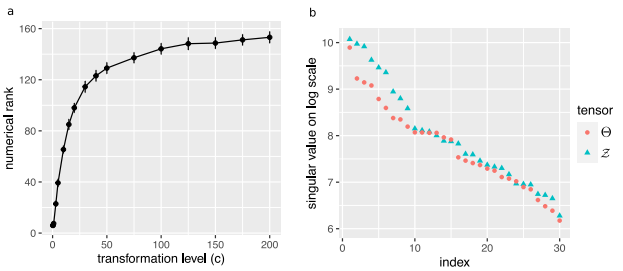

Figure 2. (a) Numerical rank of $\Theta = f(\mathcal{Z})$ versus c in Example 1. Here, the numerical rank is computed as the minimal rank for which the relative least-squares error is below 0.1, and $\mathcal{Z}$ is a rank-3 tensor with i.i.d. $N(0, 1)$ entries in the (unnormalized) singular vectors. Reported ranks are averaged across 10 replicates of $\mathcal{Z}$, with standard errors given in error bars. (b) Top $d = 30$ tensor singular values in Example 2. Numerical values in both figures are obtained by running CP decomposition with random initialization.

In the above and many other examples, the signal tensors Θ exhibit high rank in spite of the special structures they admit. These structures are hardly detectable using classical low-rank models. New methods that allow flexible tensor modeling have yet to be developed.

3.2. Overview of our proposal

Before describing our main results, we provide the intuition behind our method. In the earlier two examples, the high-rankness in the signal Θ makes the efficient estimation challenging. However, if we examine the sign of the π-shifted signal, $\mathrm{sgn}(\Theta - \pi)$, for an arbitrary π in the range of observations, then these sign tensors exhibit easily detectable low-rankness. For instance, the signal tensor in example 1 has the same sign pattern as a rank-4 tensor, since $\mathrm{sgn}(\Theta - \pi) = \mathrm{sgn}(\mathcal{Z} - f^{-1}(\pi))$. The signal tensor in example 2 has the same sign pattern as a rank-2 tensor, since $\mathrm{sgn}(\Theta - \pi) = \mathrm{sgn}(\max(i, j, k) - d(e^\pi - 1))$.

The above observation suggests a general framework to analyze both low- and high-rank signal tensors. Figure 1 illustrates the main crux of our method. We dichotomize the data tensor into a series of sign tensors, $\mathrm{sgn}(\mathcal{Y}_\Omega - \pi)$, for $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$, and then estimate the sign signals, $\mathrm{sgn}(\Theta - \pi)$, by performing classification

$$\hat{\mathcal{Z}}_\pi = \underset{\text{low rank tensor } \mathcal{Z}}{\arg\min} L(\mathrm{sgn}(\mathcal{Z}),\ \mathrm{sgn}(\mathcal{Y}_\Omega - \pi)).$$

Here, $L(\cdot, \cdot)$ denotes a carefully-designed classification objective function which will be described in later sections. The final proposed tensor estimate is

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

Our approach is built on the nonparametric sign representation of signal tensors, and the estimate $\hat{\Theta}$ is essentially estimated from dichotomized tensor series $\{\mathrm{sgn}(\mathcal{Y}_\Omega - \pi) \colon \pi \in \mathcal{H}\}$. Surprisingly, we show that proper analysis based on dichotomized data not only preserves all information in the original signals, but also brings benefits of accuracy and flexibility over classical low-rank models. The method enjoys both statistical effectiveness and computational efficiency.

4. Sign representable tensors

In this section, we develop sign representable tensor models. The algebraic and statistical characterization of sign tensor series provides the accuracy guarantee for our method.

4.1. Sign-rank and sign tensor series

Let Θ be a continuous-valued tensor, and $\mathrm{sgn}(\Theta)$ be the corresponding sign patten. The sign pattern induces an equivalence relationship between tensors. Two tensors are called sign equivalent, denoted $\simeq$, if they share the same sign patterns.

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

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

The sign-rank is also called *support rank* (Cohn and Umans, 2013), *minimal rank* (Alon et al., 2016), and *nondeterministic rank* (De Wolf, 2003) in the filed of combinatorics and information theory. Earlier work defines sign-rank for binary tensors/matrices only; we extend the notion to continuous-valued tensors. Note that the sign-rank concerns only the sign pattern but discards the magnitudes information of Θ. In particular, $\mathrm{srank}(\Theta) = \mathrm{srank}(\mathrm{sgn}\Theta)$.

Like most tensor problems (Hillar and Lim, 2013), determining the sign-rank for a general tensor is NP hard (Alon et al., 2016). Fortunately, tensors arisen in application often possess special structures that facilitate analysis. We show that the family of low sign-rank tensors is strictly broader than usual low-rank tensors. This is because sign-rank is upper bounded by the usual rank. More generally,

Corollary 1 (Upper bounds of sign-rank)**.** For any monotonic function $g \colon \mathbb{R} \to \mathbb{R}$ with $g(0) = 0$,

$$\mathrm{srank}(\Theta) \le \mathrm{rank}(g(\Theta)).$$

Conversely, the sign-rank can be dramatically smaller than the usual rank, as we have shown in Section 3.1.

Proposition 1 (Broadness)**.** *For every order $K \ge 2$ and dimension d, there exist order-K $(d, \ldots, d)$-dimensional tensors Θ such that* $\mathrm{rank}(\Theta) = d$ *and* $\mathrm{srank}(\Theta) = 2$.

We provide several common examples in Appendix where the tensor rank grows with dimension d but the sign-rank remains a constant. The results highlight the advantages of using sign-rank in the high-dimensional tensor analysis. Corollary 1 and Proposition 1 together demonstrate the strict broadness of low sign-rank tensor family over the usual low-rank tensor family.

We are now ready to introduce a family of tensors, which we coin as "sign representable tensors", as the signal tensors in model (2). Without loss of generality, assume tensor entries are bounded by $A = 1$.

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

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

Example 1 (CP/Tucker low-rank models)**.** The CP and Tucker low-rank models are the two most popular tensor

models (Anandkumar et al., 2014; Montanari and Sun, 2018; Kadmon and Ganguli, 2018; Cai et al., 2019). Let Θ be a low-rank tensor with CP rank r. We see that Θ belongs to the sign representable family, i.e., $\Theta \in \mathscr{P}_{\text{sgn}}(r+1)$ (the constant 1 is due to $\text{rank}(\Theta - \pi) \le r+1$). Similarly, Tucker low-rank tensors $\Theta \in \mathscr{P}_{\text{sgn}}(r+1)$, where $r = \prod_k r_k$ with r_k being the k-th mode Tucker rank of Θ.

Example 2 (Generalized linear models (GLMs))**.** Let $\mathcal{Y}$ be a binary tensor from a tensor logistic model (Wang and Li, 2020) with mean $\Theta = \mathbb{E}(\mathcal{Y}) = \text{logit}(\mathcal{Z})$, where $\mathcal{Z}$ is a latent low-rank tensor. Then, Θ is a special (parametric) case of sign representable tensors. Same conclusion holds for general exponential-family tensors with (known) distribution-specific link functions (Hong et al., 2020).

Example 3 (Single index models (SIMs))**.** Single index model is a flexible semiparametric model initially proposed in economics (Robinson, 1988) and has recently been popular in high-dimensional statistics (Balabdaoui et al., 2019; Ganti et al., 2017; Alquier and Biau, 2013). We here extend the model to high-dimensional tensors Θ. The SIM assumes the existence of a (unknown) monotonic function $g\colon \mathbb{R} \to \mathbb{R}$ such that $g(\Theta)$ has rank r. We see that Θ belongs to the sign representable family; i.e., $\Theta \in \mathscr{P}_{\text{sgn}}(r+1)$.

Example 4 (Tensor block models (TBMs))**.** Tensor block model (Wang and Zeng, 2019; Chi et al., 2020) assumes a checkerbord structure among tensor entries under marginal index permutation. The signal tensor Θ takes at most r distinct values, where r is the total number of multiway blocks. Our model incorporates TBM because $\Theta \in \mathscr{P}_{\text{sgn}}(r)$.

Example 5 (Min/Max hypergraphon)**.** Graphon is a popular nonparametric model for networks (Chan and Airoldi, 2014; Xu, 2018), and we have extended the model for tensors in Section 3.1. Here we revisit the model for generality. Let Θ be an order-K tensor generated from the hypergraphon $\Theta(i_1, \dots, i_K) = \log(1 + \max_k x^{(k)}_{i_k})$, where $x^{(k)}_{i_k} \sim \text{Unif}[0,1]$ i.i.d. for all $i_k \in [d_k]$ and $k \in [K]$. Every sign tensor $\text{sgn}(\Theta - \pi)$ in the series of $\pi \in [0,\ \log 2]$ is a block tensor with at most two blocks, so $\Theta \in \mathscr{P}_{\text{sgn}}(2)$.

More generally, let $g(\cdot)$ be a continuous univariate function with at most $r \ge 1$ real-valued roots in the equation $g(z) = \pi$; this property holds, e.g., when $g(z)$ is a polynomial of degree r. Then, the tensor Θ generated from $\Theta(i_1, \dots, i_K) = g(\max_k x^{(k)}_{i_k})$ belongs to $\mathscr{P}_{\text{sgn}}(r+1)$. Same conclusion applies if the maximum is replaced by minimum.

4.2. Statistical characterization of sign tensors via weighted classification

Accurate estimation of a sign representable tensor crucially depends on the behavior of sign tensor series $\text{sgn}(\Theta - \pi)$. In this section, we show that weighted classification completely characterizes the sign tensors. The results bridge the algebraic and statistical properties of sign representable tensors, thereby providing the theoretical guarantee for our nonparametric algorithm (Figure 1).

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 all $\omega \in \Omega$. We propose a weighted classification objective

$$L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \underbrace{|\bar{\mathcal{Y}}(\omega)|}_{\text{entry-specific weight}} \times \underbrace{|\text{sgn}\mathcal{Z}(\omega) - \text{sgn}\bar{\mathcal{Y}}(\omega)|}_{\text{classification loss}}, \tag{3}$$

where $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the decision variable to be optimized, $|\bar{\mathcal{Y}}(\omega)|$ is the entry-specific weight equal to the distance from the tensor entry to the target level π. The objective reduces to usual classification loss in the special case when the data tensor is binary and the target level $\pi = 0$.

Our proposed entry-specific weight is important for characterizing $\text{sgn}(\Theta - \pi)$, as we show now. Define the weighted classification risk

$$\text{Risk}(\mathcal{Z}) = \mathbb{E}_{\mathcal{Y}_\Omega} L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega), \tag{4}$$

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

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

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

The results show that the sign tensor $\text{sgn}(\Theta - \pi)$ optimizes the weighted classification risk. This fact suggests a practical procedure to estimate $\text{sgn}(\Theta - \pi)$ via optimizing the empirical risk $L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega)$. The entry-specific weight incorporates the magnitude information in the classification, where entries far away from the target level are penalized more heavily in the objective.

In order to establish the recovery guarantee, we address the uniqueness (up to sign equivalence) of the optimizer for $\text{Risk}(\cdot)$. Essentially, this is a converse problem of Proposition 2 asking whether optimizing $\text{Risk}(\cdot)$ is sufficient for recovering $\text{sgn}(\Theta - \pi)$. The local behavior of Θ around π turn out to play a key role in the accuracy guarantee.

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

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

$$\sup_{0 \leq t < \rho(\pi, \mathcal{N})} \frac{\mathbb{P}_{\omega \sim \Pi}(\omega \colon |\Theta(\omega) - \pi| \leq t)}{t^{\alpha}} \leq c, \qquad (5)$$

where $\rho(\pi, \mathcal{N}) := \min_{\pi' \in \mathcal{N}} |\pi - \pi'|$ denotes the distance from π to the nearest point in $\mathcal{N}$. The largest possible $\alpha = \alpha(\pi)$ is called the smoothness index at level π. We make the convention that $\alpha = \infty$ if the set $\{\omega \colon |\Theta(\omega) - \pi| \leq t\}$ has zero measure, implying few entries of Θ around the level π. We call the tensor Θ is α-globally smooth, if (5) holds with a global constant $c > 0$ for all $\pi \in [-1, 1]$ except for a finite number of levels.

The smoothness index α quantifies the intrinsic hardness of recovering $\text{sgn}(\Theta - \pi)$ from $\text{Risk}(\cdot)$. The recovery is more difficult at levels where the point mass concentrates (small α). The value of α is determined by the marginal density of $\Theta(\omega)$ and the sampling distribution $\omega \sim \Pi$. A higher value of $\alpha > 1$ corresponds a plateau-type zero density of $\Theta(\omega)$ around π, whereas a lower value of $\alpha < 1$ indicates the nonexistence (infinite) density at level π. A typical case is $\alpha = 1$ when $\Theta(\omega)$ has positive density bounded away from zero in the vicinity of π.

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

$$\text{MAE}(\Theta_1, \Theta_2) \stackrel{\text{def}}{=} \mathbb{E}_{\omega \sim \Pi} \|\Theta_1 - \Theta_2\|_1.$$

Theorem 4.1 (Identifiability)**.** *Under Assumption 1, for tensors $\bar{\Theta} \simeq \text{sgn}(\Theta - \pi)$ and all tensors $\mathcal{Z} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$,*

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

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

The result immediately suggests the uniqueness of the optimizer for $\text{Risk}(\cdot)$ up to a zero-measure set under Π, as long as $\alpha \neq 0$. Furthermore, the bound establishes the stability of recovering sign tensors $\text{sgn}(\Theta - \pi)$ from optimizing the population risk (4). We find that a higher value of α implies more stable recovery. Similar results hold for the optimization problem with the empirical risk (3) (Section 5).

We conclude this section by relating Assumption 1 to the examples described in Section 4.1. We see that the tensor block model is ∞-globally smooth. This is because there are finitely many elements in $\mathcal{N}$, i.e, the set of distinct block means in Θ. Furthermore, we have $\alpha = \infty$ for all $\pi \notin \mathcal{N}$, since the numerator in (5) is zero for all such π. The min/max hypergaphon with a r-degree polynomial function is 1-globally smooth, because $\alpha = 1$ for all π in the function range except at most $(r-1)$ many (stationary) levels.

5. Nonparametric tensor completion via sign series

In this section, we establish the statistical error bound for our nonparametric tensor method (Figure 1).

5.1. Estimation error and sample complexity

We cast the tensor completion problem into a series of weighted classifications, and propose the estimator

$$\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \text{sgn}\hat{\mathcal{Z}}_{\pi}, \qquad (6)$$

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

$$\hat{\mathcal{Z}}_{\pi} = \underset{\mathcal{Z} \colon \text{rank}(\mathcal{Z}) \leq r,\ \|\mathcal{Z}\|_F \leq c}{\arg\min} L(\mathcal{Z}, \text{sgn}(\mathcal{Y}_{\Omega} - \pi)), \qquad (7)$$

Here $L(\cdot, \cdot)$ denotes the weighted classification objective (3), where we have plugged in the π-shifted data tensor in the expression, and $c > 0$ is a prespecified constant independent of dimension. The rank constraint on $\mathcal{Z}$ is based on Proposition 2, and the Frobenius-norm constraint is a technical condition to simplify the analysis; the constraint does not alter the solution because $L(\mathcal{Z}, \cdot) = L(\text{sgn}\mathcal{Z}, \cdot)$

The next theorem establish the statistical convergence for the estimated sign tensor series (7). For simplicity, we assume $d_1 = \ldots = d_K = d$.

Theorem 5.1 (Estimation of sign series)**.** *Suppose $\Theta \in \mathscr{P}_{\text{sgn}}(r)$ and Assumption 1 holds at all level $\pi \notin \mathcal{N}$ with smoothness index $\alpha \in [0, 1]$. Then, for every such π, with very high probability over $\mathcal{Y}_{\Omega}$,*

$$\text{MAE}(\text{sgn}\hat{\mathcal{Z}}_{\pi}, \text{sgn}(\Theta - \pi)) \lesssim \left(\frac{dr}{|\Omega|}\right)^{\alpha/(\alpha+2)} + \frac{1}{\rho^2(\pi, \mathcal{N})} \frac{dr}{|\Omega|}.$$

Theorem 5.1 provides the error bound for the estimated sign tensor series. The results shows that our sign estimator achieves consistent recovery using as few as $\tilde{O}(dr)$ noisy observations. Furthermore, our sign tensor estimation reaches a fast rate $d^{-(K-1)}$ when $\alpha = \infty$. The convergence rate becomes favorable as the order of data tensor increases.

Using the connection between sign representable tensors and sign series, we obtain the main results on our nonparametric tensor estimation.

Theorem 5.2 (Tensor estimation error)**.** *Consider the same condition of Theorem 5.1 and the nonparametric estimate $\hat{\Theta}$ in (6). With very high probability over $\mathcal{Y}_{\Omega}$,*

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

Setting $H \asymp \left(\frac{|\Omega|}{rd}\right)^{1/2}$ *gives the estimation error bound*

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

Theorem 5.2 demonstrates the fast convergence rate of our nonparamatric tensor estimation. Our results reveal three sources of errors: the estimation error inherited from sign estimation, the bias and the variance from sign series representations. The resolution parameter H controls the bias-variance tradeoff.

Take-away: consistent is achieved whenever $|\Omega|/dr \to \infty$. Furthermore, with full observation $|\Omega| = d^K$,

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

no faster than $d^{-(K-1)/2}$. Assuming MAE $\sim$ RMSE, the same (minimax) rate as in my NIPS paper.

Corollary 2 (Sample complexity for completion)**.** With very high probability

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

Example 6 (Tensor block model)**.** Consider tensor block model. $\alpha = \infty$. Earlier results show the estimation error

$$\mathrm{MAE}(\hat{\Theta}, \Theta) \leq \sqrt{\frac{r \log d}{|\Omega|}}.$$

Compare to existing literature. Assuming MAE $\sim$ RMSE, the same (minimax) rate as in my NIPS paper.

Example 7 (Single index and Min/Max hypographon tensor model)**.** $\mathcal{O}(d^{-(K-1)/3})$. Slightly worst than Ganti et al. (2015).

5.2. Numerical implementation

In practice, we solve the optimization (7) using an efficient divide-and-conquer fashion. The estimation of sign tensors is parallelable for the series $\pi \in \mathcal{H}$ through parallel implementation.

For each $\pi \in \mathcal{H}$, an alternating optimization algorithm is developed to solve (7). Following the common practice in classification, we replace the binary loss $\ell(z, y) = |\mathrm{sgn}z - \mathrm{sgn}y|$ with a more manageable surrogate loss $F(m)$ as a function of margin $m \overset{\text{def}}{=} z\mathrm{sgn}(y)$. Examples of large-margin loss are hinge loss $F(m) = (1-m)_+$ for support vector machines, logistic loss $F(m) = \log(1+e^{-m})$ for important vector machines, and nonconvex ψ-loss $F(m) = 2\min(1, (1-m)_+)$ with $m_+ = \max(m, 0)$. We implement the hinge loss and logistic loss for illustration, although our framework is applicable to general large-margin losses (Bartlett et al., 2006).

The rank constraint is implemented by using the rank decomposition as in (1) and optimizing one factor of $\mathcal{Z}$ at a time while holding others fixed. Each suboptimization reduces to a convex optimization with a (rd_k)-dimensional decision variable. As common in tensor optimization, we run the algorithm from multiple initializations for locate a final estimate with the lowest objective value. The full procedure is described in Algorithm 1.

Algorithm 1 Nonparametric tensor completion

Input: Noisy and incomplete data tensor $\mathcal{Y}_\Omega$, rank r.
Output: Estimated signal tensor $\hat{\Theta} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$.

1: **for** $\pi \in \mathcal{H} = \{-1, \ldots, -\frac{1}{H}, 0, \frac{1}{H}, \ldots, 1\}$ **do**
2: Estimate sign tensor $\mathrm{sgn}(\mathcal{Z}_\pi)$ by performing weighted classification using sub-algorithm.
3: **end for**
4: Return estimated tensor $\hat{\Theta} = \frac{1}{2H+1} \sum_{\pi \in \mathcal{H}} \mathrm{sgn}(\mathcal{Z}_\pi)$.

Sub-algorithm: Sign tensor estimation using weighted classification

Input: Noisy and incomplete data tensor $\mathcal{Y}_\Omega$, rank r, target level π.
Output: Sign tensor $\mathrm{sgn}(\mathcal{Z}) \in \{-1, 1\}^{d_1 \times \cdots \times d_K}$ as the estimation of $\mathrm{sgn}(\Theta - \pi)$.

5: Random initialization of tensor factors $\boldsymbol{A}_k = [\boldsymbol{a}_1^{(k)}, \ldots, \boldsymbol{a}_r^{(k)}] \in \mathbb{R}^{d_k \times r}$ for all $k \in [K]$.
6: Normalize columns of $\boldsymbol{A}_k$ to have unit-norm for $k \in [K-1]$, and absorb the scales into the columns of $\boldsymbol{A}_K$.
7: **while** not convergence **do**
8: **for** $k = 1, \ldots, K$ **do**
9: Update $\boldsymbol{A}_k$ while holding others fixed: $\boldsymbol{A}_k \leftarrow \arg\min_{\boldsymbol{A}_k \in \mathbb{R}^{d_k \times r}} \sum_{\omega \in \Omega} |\mathcal{Y}(\omega) - \pi| F(\mathcal{Z}(\omega)\mathrm{sgn}(\mathcal{Y}(\omega) - \pi))$, where $F(\cdot)$ is the large-margin loss, and $\mathcal{Z} = \sum_{s \in [r]} \boldsymbol{a}_s^{(1)} \otimes \cdots \otimes \boldsymbol{a}_s^{(K)}$.
10: **end for**
11: **end while**

Example 8 (Addition examples that satisfying Proposition 1)**.** We provide a tensor example with $\text{rank}(\Theta) = d$ but $\text{srank}(\Theta) = 3$. Define $\Theta = \sum_{r=1}^{d} \boldsymbol{e}_r^{\otimes 2} \otimes \mathbf{1}_d^{\otimes(K-2)}$, where $\boldsymbol{e}_r = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is the r-th canonical basis in $\mathbb{R}^d$, and $\mathbf{1}_d \in \mathbb{R}^d$ is a vector with all entries 1. Based on the definition of Θ, we have

$$\text{rank}(\Theta) = \text{rank}(\boldsymbol{I}), \quad \text{srank}(\Theta) = \text{srank}(\boldsymbol{I}),$$

where $\boldsymbol{I} \in \mathbb{R}^{d \times d}$ is the identity matrix. Therefore, it suffices to show that $\text{srank}(\boldsymbol{I}) = 3$. We now construct a rank-2 matrix $\boldsymbol{A}$ such that $\text{sgn}(\boldsymbol{A} - 1/2) = \text{sgn}(\boldsymbol{I})$. Define

$$\boldsymbol{A} = \begin{bmatrix} 1 & -\frac{1}{2} \times 1 \\ 2^{-1} & -\frac{1}{2} \times 4^{-1} \\ \vdots & \vdots \\ 2^{-d+1} & -\frac{1}{2} \times 4^{-d+1} \end{bmatrix} \begin{bmatrix} 1 & 2 & \cdots & 2^{d-1} \\ 1 & 4 & \cdots & 4^{d-1} \end{bmatrix}.$$

It is easy to verify that $\boldsymbol{A}(i,j) = \frac{1}{2}$ if $i = j$, and $\boldsymbol{A}(i,j) < \frac{1}{2}$ otherwise. Therefore, $\text{sgn}(\boldsymbol{A} - 1/2) = \boldsymbol{I}$.

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

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) = \mathbb{E}_{\omega \sim \Pi} \underbrace{\mathbb{E}_{\mathcal{Y}(\omega)} \left\{ |\mathcal{Y}(\omega) - \pi| \left[\left| \mathcal{Z}(\omega) - \text{sgn}(\bar{\mathcal{Y}}(\omega)) \right| - \left| \bar{\Theta}(\omega) - \text{sgn}(\bar{\mathcal{Y}}(\omega)) \right| \right] \right\}}_{=: I(\omega)}. \tag{8}$$

Denote $y = \mathcal{Y}(\omega)$, $z = \mathcal{Z}(\omega)$, $\bar{\theta} = \bar{\Theta}(\omega)$, and $\theta = \Theta(\omega)$. It follows from the expression of $I(\omega)$ that

$$\begin{aligned} I(\omega) &= \mathbb{E}_y \left[(y - \pi)(\bar{\theta} - z) \mathbb{1}(y \geq \pi) + (\pi - y)(z - \bar{\theta}) \mathbb{1}(y < \pi) \right] \\ &= \mathbb{E}_y \left[(\bar{\theta} - z)(y - \pi) \right] \\ &= \left[\text{sgn}(\theta - \pi) - z \right] (\theta - \pi) \\ &= |\text{sgn}(\theta - \pi) - z| |\theta - \pi| \geq 0 \end{aligned} \tag{9}$$

where the third line uses the fact $\mathbb{E}y = \theta$ and $\bar{\theta} = \text{sgn}(\theta - \pi)$, and the last line uses the assumption $z \in \{-1, 1\}$. In particular, the equality is attained when $z = \text{sgn}(\theta - \pi)$ or $\theta = \pi$. Combining (9) with (8), we conclude

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

for all $\mathcal{Z} \in \{-1, 1\}^{d_1 \times \cdots \times d_K}$. Therefore,

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

Because $\text{srank}(\bar{\Theta}) \le r$ by assumption, the last inequality becomes equality. The proof is complete. □

Proof. We verify two conditions.

1. Approximation error. For $\mathcal{Z}$ with $\text{rank}(\mathcal{Z}) \le r$, we have $\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) = 0$ for all d.
2. Variance-to-mean relationship

$$\text{Var}_{\mathcal{Y},\Omega}[L(\mathcal{Z}, \bar{\mathcal{Y}}_\pi) - L(\bar{\Theta}, \mathcal{Y}_\pi)] \le [\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})]^{\alpha/(1+\alpha)} + \frac{1}{\rho(\pi, \mathcal{N})}[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})].$$

Apply Lemma 1 to the above condition, we obtain

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) \le t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho(\pi, \mathcal{N})} t_n, \quad \text{where } t_n = \frac{Krd}{n}.$$

□

Lemma 1. *Because the classification rate is scale-free; Risk*$(\mathcal{Z}) =$ *Risk*$(c\mathcal{Z})$ *for every* $c > 0$*. Therefore, without loss of generality, we solve the estimate subject to* $\|\mathcal{Z}\|_F \le 1$*,*

$$\hat{\mathcal{Z}} = \underset{\mathcal{Z}\colon \text{rank}(\mathcal{Z}) \le r, \|\mathcal{Z}\|_F \le 1}{\arg\min} L(\mathcal{Z}, \bar{\mathcal{Y}}_\pi).$$

Write $|\Omega| = n$*. We have*

$$\mathbb{P}[\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta}) \ge t_n] \le \frac{7}{2}\exp(-Cnt_n).$$

The rate of convergence $t_n > 0$ *is determined by the solution to the following inequality,*

$$\frac{1}{t_n}\int_{t_n}^{\sqrt{t_n^\alpha + \rho^{-1}t_n}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon,\ \mathcal{F},\ \|\cdot\|_2)}d\varepsilon \le n^{1/2},$$

where $\mathcal{F} = \{\mathcal{Z}\colon \text{rank}(\mathcal{Z}) \le r,\ \|\mathcal{Z}\|_F^2 \le 1\}$ *and* $\rho = \rho(\pi, \mathcal{N})$*. By Lemma 2, we obtain*

$$t_n \asymp \left(\frac{Kdr}{n}\right)^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho^2(\pi, \mathcal{N})}\frac{Kdr}{n}.$$

Finally, we obtain

$$\mathbb{P}[\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta}) \ge t_n] \le \frac{7}{2}\exp(-Cd^{\frac{\alpha+1}{\alpha+2}}n^{\frac{1}{\alpha+2}}) \le \frac{7}{2}\exp(-C\sqrt{d}),$$

where $C = C(k, r) > 0$ *is a constant independent of* d *and* n*.*

Lemma 2 (Bracketing number for bounded low rank tensor)**.**

$$\sqrt{\mathbb{E}_{\omega\sim\Pi}|\mathcal{Z}_1(\omega) - \mathcal{Z}_2(\omega)|^2} \le \|\mathcal{Z}_1 - \mathcal{Z}_2\|_\infty \le \|\mathcal{Z}_1 - \mathcal{Z}_2\|_F.$$

Therefore

$$\mathcal{H}_{[\,]}(2\varepsilon, \mathcal{F}, \|\cdot\|_2) \le \mathcal{H}(\varepsilon, \mathcal{F}, \|\cdot\|_F) \le C(1 + Kdr)\log\frac{d}{\varepsilon},$$

where the covering number for low rank tensor is based on Mu et al. (2014); Ibrahim et al. (2020).

References

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

Alquier, P. and G. Biau (2013). Sparse single-index model. *Journal of Machine Learning Research 14*(Jan), 243–280.

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

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

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

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

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

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

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

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

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

Ganti, R. S., L. Balzano, and R. Willett (2015). Matrix completion under monotonic single index models. *Advances in Neural Information Processing Systems 28*, 1873–1881.

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

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

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

Ibrahim, S., X. Fu, and X. Li (2020). On recoverability of randomly compressed tensors with low cp rank. *IEEE Signal Processing Letters 27*, 1125–1129.

Kadmon, J. and S. Ganguli (2018). Statistical mechanics of low-rank tensor decomposition. *Advances in Neural Information Processing Systems 31*, 8201–8212.

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

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

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

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

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

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