Appendix for “Beyond the Signs: Nonparametric Tensor Completion via Sign Series”

The appendix consists of additional theoretical results (Section A), numerical experiments (Section B), and proofs (Section C).

A Additional results

A.1 Sensitivity of tensor rank to monotonic transformation

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

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

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

where $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{30}$ are vectors consisting of $N(0,1)$ entries. Then we 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 of Θ is determined as the minimal rank for which the relative approximation error is below 0.1, i.e.,

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

We compute $\hat{r}(\Theta)$ by searching over $s \in \{1, \ldots, 30^2\}$, where for each s, we (approximately) solve the least-square minimization using CP function in R package `rTensor`.

We repeat steps 1-2 ten times, and report the averaged numerical rank of Θ along with transformation level c in Figure 1a.

A.2 Tensor rank and sign-rank

In the main paper, we have provided several tensor examples with high usual rank but low sign-rank. This section provides more examples and their proofs.

Unless otherwise specified, let Θ be an order-K $(d, \ldots, d)$-dimensional tensor.

Example A.1 (Max hypergraphon)**.** Suppose the tensor Θ takes the form

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

Then

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

Proof. We first prove the results for $K = 2$. We can check that Θ has a full rank i.e. $\text{rank}(\Theta) = d$ for $K = 2$ from elementary row operations as follows.

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

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

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

Therefore, the matrix $\text{sgn}(\Theta - \pi)$ is a rank-2 block matrix. By definition of sign rank, we have $\text{srank}(\Theta - \pi) = 2$. We now extend the results to $K \geq 3$. Based on the definition, the rank of a tensor is lower bounded by the rank of its matrix slice. It follows that $\text{rank}(\Theta) \geq \text{rank}(\Theta(:\,,:\,,1,\ldots,1)) = d$. For the sign rank with feasible π, notice that the sign tensor $\text{sgn}(\Theta - \pi)$ takes the similar form as in (1),

$$\text{sgn}(\Theta(i_1, \ldots, i_K) - \pi) = \begin{cases} -1, & i_k < i^* \text{ for all } k \in [K]; \\ 1, & \text{otherwise,} \end{cases} \tag{2}$$

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

In fact, the example A.1 is a special case of the following Proposition.

Proposition A.1 (Min/Max hypergraphon)**.** *Let $\mathcal{Z}_{\max} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote a tensor with entries*

$$\mathcal{Z}_{\max}(i_1, \ldots, i_K) = \max(x^{(1)}_{i_1}, \ldots, x^{(K)}_{i_K}), \tag{3}$$

where $x^{(1)}_{i_1}, \ldots, x^{(K)}_{i_K} \in [0,1]$ are pre-specified numbers for all $i_k \in [d_k]$ and $k \in [K]$. Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuous function and $\Theta := g(\mathcal{Z}_{\max})$ be the transformed tensor. Suppose the function $g(z) = \pi$ for a fixed $\pi \in \mathbb{R}$ has at most $r \geq 1$ distinct real roots, then sign rank of $(\Theta - \pi)$ satisfies

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

The same conclusion holds if we use min *in place of* max *in* (3)*.*

Proof. We reorder the tensor index along each mode such that $x^{(k)}_1 \leq \cdots \leq x^{(k)}_{d_k}$ 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 A.1, 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 we have denoted the binary vectors

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

  Therefore, $\operatorname{srank}(\Theta - \pi) \leq \operatorname{rank}(\operatorname{sgn}(\Theta - \pi)) = 2$.

- When $r \geq 2$. Based on the 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 $\operatorname{sgn}(\Theta - \pi) = -1$,

$$\begin{aligned}\{\omega \colon \Theta(\omega) - \pi < 0\} &= \{\omega \colon g(\mathcal{Z}_{\max}(\omega)) < \pi\} \\ &= \cup_{s \in \mathcal{I}} \{\omega \colon \mathcal{Z}_{\max}(\omega) \in (z_s, z_{s+1})\} \\ &= \cup_{s \in \mathcal{I}} \{\omega \colon x_{i_k}^{(k)} < z_{s+1} \text{ for all } k \in [K]\} \cap \{\omega \colon x_{i_k}^{(k)} \leq z_s \text{ for all } k \in [K]\}^c (4)\end{aligned}$$

  The identity (4) is equivalent to

$$\mathbb{1}(\Theta(i_1, \ldots, i_K) < \pi) = \sum_{s \in \mathcal{I}} \left(\prod_k \mathbb{1}(x_{i_k}^{(k)} < z_{s+1}) - \prod_k \mathbb{1}(x_{i_k}^{(k)} \leq z_s) \right), \tag{5}$$

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

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

  where we have denoted the two binary vectors

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

  Note that $|\mathcal{I}| \leq r - 1$. Therefore we conclude that

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

Combining two cases yield that $\operatorname{srank}(\Theta - \pi) \leq 2r$ for any $r \geq 1$. □

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

Proposition A.2. *Let $\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2}$ be a matrix. For any given $K \geq 3$, define an order-K tensor $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ by*

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

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

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

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

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

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

Proof. 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 directly shows that $\boldsymbol{M}$ is full rank as follows.

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

We now show $\operatorname{srank}(\boldsymbol{M} - \pi) \leq 3$ 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

$$\boldsymbol{A} = \boldsymbol{b} \otimes \operatorname{rev}(\boldsymbol{b}) + \operatorname{rev}(\boldsymbol{b}) \otimes \boldsymbol{b}. \tag{6}$$

It is easy to see that $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is a banded matrix 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 as $|i-j|$ gets close to 0 i.e.,

$$\boldsymbol{A}(i,j) \geq \boldsymbol{A}(i',j'), \quad \text{for all } |i-j| \geq |i'-j'|. \tag{7}$$

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 (7). By definition (6), $\boldsymbol{A}$ is a rank-2 matrix. Henceforce, $\operatorname{srank}(\boldsymbol{M} - \pi) = \operatorname{srank}(\boldsymbol{A} - \pi') \leq 3$. □

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

Example A.3 (Stacked identity matrices)**.** Let $\boldsymbol{I}$ be a d-by-d diagonal matrix. Then

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

Proof. Depending on the value of π, the sign matrix $\operatorname{sgn}(\boldsymbol{I} - \pi)$ fall into one of the three cases: 1) $\operatorname{sgn}(\boldsymbol{I} - \pi)$ is a matrix of all 1; 2) $\operatorname{sgn}(\boldsymbol{I} - \pi)$ is a matrix of all -1; 3) $\operatorname{sgn}(\boldsymbol{I} - \pi) = \operatorname{sgn}(\boldsymbol{I})$. The former two cases are trivial, so it suffices to show $\operatorname{srank}(\boldsymbol{I}) \leq 3$.

Based on Example A.2, the rank-2 matrix $\boldsymbol{A}$ in (6) satisfies

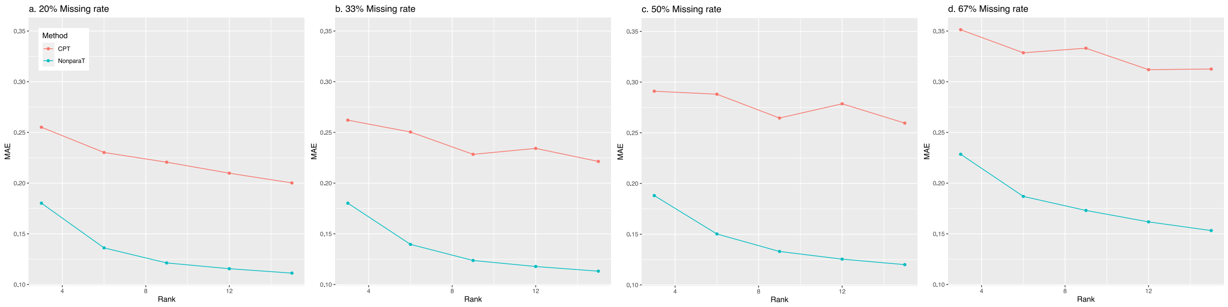
$$\boldsymbol{A}(i,j) \begin{cases} = 2^{-d}, & i = j, \\ \geq 2^{-d} + 2^{-d-2}, & i \neq j. \end{cases}$$

Therefore, $\operatorname{sgn}\left(2^{-d} + 2^{-d-3} - \boldsymbol{A}\right) = \operatorname{sgn}(\boldsymbol{I})$. We conclude that $\operatorname{srank}(\boldsymbol{I}) \leq \operatorname{rank}(2^{-d} + 2^{-d-3} - \boldsymbol{A}) = 3$.

□

B Additional data analysis

B.1 Brain connectivity analysis



Supplementary Figure S1: Estimation errors versus rank on cross-validation. Four different missing rates are used: 20%,33%,50%, and 67%.

Figure B.1 shows the MAE based on cross-validation with $r = 3, 6, 9, 12, 15$ and $H = 20$. We find that our method outperforms substantially in all combinations of ranks and missing rates. The figure shows that the higher missing rates, the greater gap of performance between the two methods, which implies the greater advantage of our method when the proportion of missing data is high. In addition, we found that small standard errors of MAE of our method from five repeated cross-validation shows the more stable performance over CPT (see Table 2 in the main paper). One advantage of our method is that our estimation is always bounded in $[0, 1]$ whereas CPT estimation may fall outside the valid range $[0, 1]$.

We estimate the signal tensor Θ from the observed binary tensor $\mathcal{Y}$, $r = 10$, and $H = 20$. We regress the denoised connection strength of brain edge (i, j) across 114 individuals $(\hat{\Theta}(i, j,))$ to IQ scores. We repeated the regression for every brain edge and Top 10 significant edges for the normalized IQ scores have been plotted in the main paper.

B.2 NIPS data analysis

In the main paper we have summarized the MAE from cross validation for $r = 6, 9, 12$. Here we provide additional results for a wider range $r = 3, 6, 9, 12, 15$. Table S1 suggests that further increment of rank appears to have little effect on the performance. In addition, we also perform naive imputation where the missing values are predicted using the observed sample average. The two tensor methods outperform the naive imputation, implying the necessity of incorporating tensor structure in the analysis.

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)				

Supplementary Table S1: Prediction accuracy measured in MAE in the NIPS data analysis. The reported MAEs are averaged over five runs of cross-validation, with standard errors in parentheses. Bold numbers indicate the minimal MAE among three methods. For low-rank CPT, we use R function `rTensor` with default hyperparameters, and for our method, we set $H=20$.

C Proofs

C.1 Proofs of Propositions 1-3

Proof of Proposition 1. Based on the definition of sign rank, it suffices to show that $\Theta \simeq g(\Theta)$. The strictly monotonicity of g implies that the inverse function $g^{-1}\colon \mathbb{R}\to\mathbb{R}$ is well-defined.

Furthermore, the mapping $x \mapsto g(x)$ is sign preserving. Specifically, if $x \geq 0$, then $g(x) \geq g(0) = 0$. On the other hand, if $g(x) \geq 0 = g(0)$, then applying g^{-1} to both sides gives $x \geq 0$. Therefore, $\Theta \simeq g(\Theta)$, and $\text{srank}(\Theta) = \text{srank}(g(\Theta)) \leq \text{rank}(g(\Theta))$. □

Proof of Proposition 2. See Proposition A.2 and Example A.2 for constructive examples. □

Proof of Proposition 3. Based on the definition of the weighted classification objective function, 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 $z \in \{-1,1\}$. The equality (9) is attained when $z = \text{sgn}(\theta-\pi)$ or $\theta = \pi$. Combining (9) with (8), we conclude that, for all $\mathcal{Z} \in \{-1,1\}^{d_1\times\cdots\times d_K}$,

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) = \mathbb{E}_{\omega\sim\Pi}|\text{sgn}(\Theta(\omega)-\pi) - \mathcal{Z}(\omega)||\Theta(\omega)-\pi| \geq 0, \tag{10}$$

In particular, setting $\mathcal{Z} = \bar{\Theta} = \text{sgn}(\Theta-\pi)$ in (10) yields the minimum. Therefore,

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

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

C.2 Proof of Theorem 1

Proof of Theorem 1. Based on (10) in Proposition 3 we have

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) = \mathbb{E}\left[|\text{sgn}\mathcal{Z} - \text{sgn}\bar{\Theta}||\bar{\Theta}|\right]. \tag{11}$$

The Assumption 1 states that

$$\mathbb{P}\left(|\bar{\Theta}| \leq t\right) \leq ct^{\alpha}, \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}). \tag{12}$$

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

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

- Case 1: $\alpha > 0$.

  By (11), for all $0 \leq t < \rho(\pi, \mathcal{N})$,

$$\begin{aligned} \text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) &\geq t\mathbb{E}\left(|\text{sgn}\mathcal{Z} - \text{sgn}\hat{\Theta}|\mathbb{1}\{|\hat{\Theta}| > t\}\right) \\ &\geq 2t\mathbb{P}\left(\text{sgn}\mathcal{Z} \neq \text{sgn}\bar{\Theta} \text{ and } |\bar{\Theta}| > t\right) \\ &\geq 2t\Big\{\mathbb{P}\left(\text{sgn}\mathcal{Z} \neq \text{sgn}\bar{\Theta}\right) - \mathbb{P}\left(|\bar{\Theta}| \leq t\right)\Big\} \\ &\geq t\Big\{\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) - 2ct^{\alpha}\Big\}, \end{aligned} \tag{13}$$

  where the last line follows from the definition of MAE and (12). We maximize the lower bound (13) with respect to t, and obtain the optimal t_{opt},

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

  The corresponding lower bound of the inequality (13) becomes

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

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

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

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

- Case 2: $\alpha = \infty$. The inequality (13) now becomes

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) \geq t\text{MAE}(\text{sgn}\bar{\Theta}, \text{sgn}\mathcal{Z}), \quad \text{for all } 0 \leq t < \rho(\pi, \mathcal{N}). \tag{14}$$

  The conclusion follows by taking $t = \frac{\rho(\pi, \mathcal{N})}{2}$ in the inequality (14).

□

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

$$\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) \lesssim [\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})}\left[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})\right], \tag{15}$$

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

C.3 Proof of Theorem 2

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

Lemma C.1 (Variance-to-mean relationship). *Consider the same setup as in Theorem 2. Fix $\pi \in [-1,1]$. Denote the π-shifted signal $\bar{\Theta} = \Theta - \pi$, the π-shifted data $\bar{\mathcal{Y}} = \mathcal{Y} - \pi$, the π-weighted classification objective function*

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

and $\mathrm{Risk}(\mathcal{Z}) = \mathbb{E}L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega)$*. Under Assumption 1 of the (α,π)-smoothness of Θ, we have*

$$\mathrm{Var}[L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - L(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)] \lesssim [\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})}[\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})], \tag{16}$$

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

Proof. We expand the variance by

$$\begin{aligned} \mathrm{Var}[L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - L(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)] &\leq \mathbb{E}|L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - L(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)|^2 \\ &\leq \mathbb{E}|L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - L(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)| \\ &\leq \mathbb{E}|\mathrm{sgn}\mathcal{Z} - \mathrm{sgn}\bar{\Theta}| = \mathrm{MAE}(\mathrm{sgn}\mathcal{Z}, \mathrm{sgn}\bar{\Theta}), \end{aligned} \tag{17}$$

where the second line comes from the boundedness of classification loss $L(\cdot,\cdot)$, and the last line comes from the inequality $||\mathrm{sgn}\mathcal{Z}(\omega) - \mathrm{sgn}\bar{\mathcal{Y}}_\Omega(\omega)| - |\mathrm{sgn}\bar{\Theta}(\omega) - \mathrm{sgn}\mathcal{Y}_\Omega(\omega)|| \leq |\mathrm{sgn}\mathcal{Z}(\omega) - \mathrm{sgn}\bar{\Theta}(\omega)|$ and the fact that the classification weight $|\bar{\mathcal{Y}}(\omega)|$ is bounded for $\omega \in \Omega$. Here we ignore constant multiplication in $\leq$ because the constants will be suppressed in the asymptotical order relationship $\lesssim$ in the end. The conclusion (16) then directly follows by applying remark C.1 to (17). □

Proof of Theorem 2. Fix $\pi \in [-1,1]$. For notational simplicity, we suppress all subscripts related to π, and write $\hat{\mathcal{Z}}$ in place of $\hat{\mathcal{Z}}_\pi$. Denote $n = |\Omega|$ and $\rho = \rho(\pi, \mathcal{N})$.

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

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

We view $\mathcal{Y}_\Omega = \{\bar{\mathcal{Y}}(\omega)\colon \omega \in \Omega\}$ as a collection of n independent random variables where the randomness is induced form both $\bar{\mathcal{Y}}$ and $\omega \sim \Pi$. The tensor $\mathcal{Z}$ induces a function $f_\mathcal{Z}\colon \Omega \mapsto \mathbb{R}$ such that $f_\mathcal{Z}(\omega) = \mathcal{Z}(\omega)$ for all $\omega \in \Omega$. Then

$$\Delta_n(f_\mathcal{Z}, f_\mathcal{Y}) := L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - L(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega), \quad \text{where } n = |\Omega|,$$

is an empirical process induced by function $f_\mathcal{Z} \in \mathcal{F}_\mathcal{T} := \{f_\mathcal{Z}\colon \mathcal{Z} \in \mathcal{T}\}$ where $\mathcal{T} = \{\mathcal{Z}\colon \mathrm{rank}(\mathcal{Z}) \leq r,\ \|\mathcal{Z}\|_F \leq 1\}$.

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

$$\mathrm{Var}\Delta_n(f_\mathcal{Z}, f_\mathcal{Y}) \lesssim [\mathbb{E}\Delta_n(f_\mathcal{Z}, f_\mathcal{Y})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho}\mathbb{E}\Delta_n(f_\mathcal{Z}, f_\mathcal{Y}), \quad \text{for all } \mathcal{Z} \in \mathcal{T}.$$

Applying the second order condition to Theorem 3 of Wang et al. (2008) gives that

$$\mathbb{P}\left(\text{Risk}(\hat{\mathcal{Z}}) - \text{Risk}(\bar{\Theta}) \geq L_n\right) \leq C\exp(-nL_n^2), \tag{18}$$

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

$$\frac{1}{L_n}\int_{L_n}^{\sqrt{L_n^{\alpha/(\alpha+1)} + \frac{L_n}{\rho}}} \sqrt{\mathcal{H}_{[\ ]}(\varepsilon, \mathcal{F}_{\mathcal{T}}, \|\cdot\|_2)}d\varepsilon \leq C'\sqrt{n}, \tag{19}$$

for some constant $C' > 0$. In particular, the smallest L_n satisfying (19) yields the best upper bound of the error rate. Here $\mathcal{H}_{[\ ]}(\varepsilon, \mathcal{F}_{\mathcal{T}}, \|\cdot\|_2)$ denotes the L_2-metric, ε-bracketing number (c.f. Definition C.1) of family $\mathcal{F}_{\mathcal{T}}$.

It remains to solve for the smallest possible L_n in (19). Based on Lemma C.2, the inequality (19) is satisfied with

$$L_n \asymp t_n^{(\alpha+1)/(\alpha+2)} + \frac{1}{\rho}t_n, \quad \text{where } t_n = \frac{Krd_{\max}\log d_{\max}}{n}. \tag{20}$$

Therefore, by (18), with very high probability.

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

Inserting the above bound into (15) gives

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

where 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 $b := \rho t_n^{-1/(\alpha+2)}$. Plugging the form of t_n in (20) completes the proof. □

Definition C.1 (Bracketing number)**.** Consider a set 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

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

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

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

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

Then, the following inequality is satisfied.

$$\frac{1}{L_n}\int_{L_n}^{\sqrt{L_n^{\alpha/(\alpha+1)}+\frac{L_n}{\rho(\pi,\mathcal{N})}}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon,\mathcal{F}_{\mathcal{T}},\|\cdot\|_2)}d\varepsilon \leq C'n^{1/2}, \tag{21}$$

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

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

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

Based on Kosorok (2007, Theorem 9.22), the L_2-metric, (2ϵ)-bracketing number in $\mathcal{F}_{\mathcal{T}}$ is bounded by

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

The last inequlity is from the upper bound of the covering number for rank-r tensor (Mu et al., 2014; Ibrahim et al., 2020).

Inserting the bracketing number into (21) gives

$$g(L) = \frac{1}{L}\int_{L}^{\sqrt{L^{\alpha/(\alpha+1)}+\rho^{-1}L}} \sqrt{Krd_{\max}\log\left(\frac{d_{\max}}{\varepsilon}\right)}d\varepsilon. \tag{22}$$

By the monotonicity of the integrand in (22), we bound $g(L)$ by

$$\begin{aligned} g(L) &\leq \frac{\sqrt{Krd_{\max}}}{L}\int_{L}^{\sqrt{L^{\alpha/\alpha+1}+\rho^{-1}L}} \sqrt{\log\left(\frac{d_{\max}}{L}\right)}d\varepsilon \\ &\leq \sqrt{Krd_{\max}(\log d_{\max} - \log L)}\left(\frac{L^{\alpha/(2\alpha+2)} + \sqrt{\rho^{-1}L}}{L} - 1\right) \\ &\leq \sqrt{Krd_{\max}(\log d_{\max})}\left(\frac{1}{L^{(\alpha+2)/(2\alpha+2)}} + \frac{1}{\sqrt{\rho L}}\right), \end{aligned} \tag{23}$$

where the second line follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$. It remains to verify that $g(L_n) \leq C'n^{1/2}$ for L_n specified in (21). Plugging L_n into the last line of (23) gives

$$\begin{aligned} g(L_n) &\leq \sqrt{Krd_{\max}(\log d_{\max})}\left(\frac{1}{L_n^{(\alpha+2)/(2\alpha+2)}} + \frac{1}{\sqrt{\rho L_n}}\right) \\ &\leq \sqrt{Krd_{\max}(\log d_{\max})}\left(\left[\left(\frac{Krd_{\max}\log d_{\max}}{n}\right)^{\frac{\alpha+1}{\alpha+2}}\right]^{-\frac{\alpha+2}{2\alpha+2}} + \left[\rho\left(\frac{Krd_{\max}\log d_{\max}}{\rho n}\right)\right]^{-\frac{1}{2}}\right) \\ &\leq C'n^{1/2}, \end{aligned}$$

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

C.4 Proof of Theorem 3

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

$$\begin{aligned}\text{MAE}(\hat{\Theta},\Theta) &= \mathbb{E}\left|\frac{1}{2H+1}\sum_{\pi\in\Pi}\text{sgn}\hat{Z}_\pi - \Theta\right| \qquad (24)\\ &\leq \mathbb{E}\left|\frac{1}{2H+1}\sum_{\pi\in\Pi}\Big(\text{sgn}\hat{Z}_\pi - \text{sgn}(\Theta-\pi)\Big)\right| + \mathbb{E}\left|\frac{1}{2H+1}\sum_{\pi\in\Pi}\text{sgn}(\Theta-\pi)-\Theta\right|\\ &\leq \frac{1}{2H+1}\sum_{\pi\in\Pi}\text{MAE}(\text{sgn}\hat{Z}_\pi,\text{sgn}(\Theta-\pi)) + \frac{1}{H}.\end{aligned}$$

Write $n = |\Omega|$. We prove that

$$\frac{1}{2H+1}\sum_{\pi\in\Pi}\text{MAE}(\text{sgn}\hat{Z}_\pi,\text{sgn}(\Theta-\pi)) \lesssim t_n^{\alpha/(\alpha+2)} + \frac{1}{H} + Ht_n, \quad \text{with } t_n = \frac{Krd_{\max}\log(d_{\max})}{n}. \qquad (25)$$

Theorem 2 implies that the sign estimation accuracy depends on the closeness of $\pi \in \mathcal{H}$ to the mass points in $\mathcal{H}$. Therefore, we partition the level set $\pi \in \mathcal{H}$ based on their closeness to $\mathcal{H}$. Specifically, let $\mathcal{N}_H \overset{\text{def}}{=} \bigcup_{\pi'\in\mathcal{N}}\left(\pi' - \frac{1}{H}, \pi' + \frac{1}{H}\right)$ denote the set of levels close to the mass points. We expand (25) by

$$\begin{aligned}&\frac{1}{2H+1}\sum_{\pi\in\Pi}\text{MAE}(\text{sgn}\hat{Z}_\pi,\text{sgn}(\Theta-\pi))\\ =&\frac{1}{2H+1}\sum_{\pi\in\Pi\cap\mathcal{N}_H}\text{MAE}(\text{sgn}\hat{Z}_\pi,\text{sgn}(\Theta-\pi)) + \frac{1}{2H+1}\sum_{\pi\in\Pi\cap\mathcal{N}_H^c}\text{MAE}(\text{sgn}\hat{Z}_\pi,\text{sgn}(\Theta-\pi)). \qquad (26)\end{aligned}$$

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

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

where the last inequality follows from the Lemma C.3. Combining the bounds for the two terms in (26) completes the proof for conclusion (25). Therefore, plugging (25) into (24) yields,

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

□

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

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

Proof. Notice that all points $\pi \in \Pi \cap \mathcal{N}_H^c$ satisfy $|\pi - \pi'| > \frac{1}{H}$ for all $\pi' \in \mathcal{N}$.

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

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

References

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.

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

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.

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

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