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 transformations

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

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

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

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

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

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

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

A.2 Tensor rank and sign-rank

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

Example 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

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

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

operations as follows

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

where Θ_i denotes the i-th row of Θ. Now it suffices to show $\operatorname{srank}(\Theta-\pi)\le 2$ 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 $\operatorname{sgn}(\Theta-\pi)$ takes the form

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

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

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

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

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

In fact, 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*

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

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

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

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

Proof. We reorder the tensor indices along each mode such that $x^{(k)}_1\le\cdots\le x^{(k)}_{d_k}$ 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 $r\ge 2$.

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

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

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

- When $r \ge 2$. 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 $1 \le s \le r-1$. 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 $\operatorname{sgn}(\Theta - \pi) = -1$,

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

The equation (4) is equivalent to

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

for all $(i_1, \dots, i_K) \in [d_1] \times \cdots \times [d_K]$, where $\mathbb{1}(\cdot) \in \{0, 1\}$ denotes the indicator function. The equation (5) implies the low-rank representation of $\operatorname{sgn}(\Theta - \pi)$,

$$\operatorname{sgn}(\Theta - \pi) = 1 - 2 \sum_{s \in \mathcal{I}} \left(\boldsymbol{a}^{(1)}_{s+1} \otimes \cdots \otimes \boldsymbol{a}^{(K)}_{s+1} - \bar{\boldsymbol{a}}^{(1)}_s \otimes \cdots \otimes \bar{\boldsymbol{a}}^{(K)}_s \right), \tag{6}$$

where we have denoted the two binary vectors

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

Therefore, by (6) and the assumption $|\mathcal{I}| \le r-1$, we conclude that

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

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

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

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

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

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

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

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

Example A.2 (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

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

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,

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

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

$$\boldsymbol{A} = \boldsymbol{b} \otimes \operatorname{rev}(\boldsymbol{b}) + \operatorname{rev}(\boldsymbol{b}) \otimes \boldsymbol{b}. \tag{7}$$

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.,

$$\boldsymbol{A}(i, j) \geq \boldsymbol{A}(i', j'), \quad \text{for all } |i - j| \geq |i' - j'|. \tag{8}$$

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 (8). By definition (7), $\boldsymbol{A}$ 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 identity 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 $\operatorname{sgn}(\boldsymbol{I} - \pi)$ falls 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) $\operatorname{sgn}(\boldsymbol{I} - \pi) = 2\boldsymbol{I} - \mathbf{1}_d \otimes \mathbf{1}_d$. The former two cases are trivial, so it suffices to show $\operatorname{srank}(\boldsymbol{I} - \pi) \leq 3$ in the third case.

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

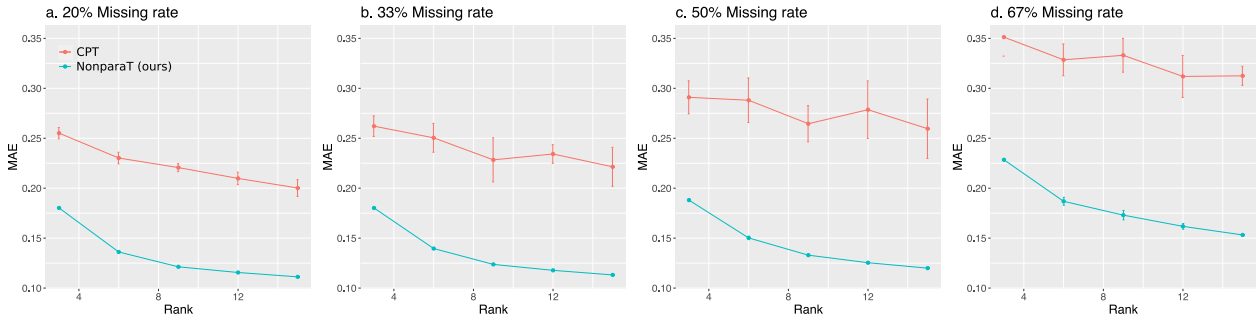
$$\boldsymbol{A}(i, j) \begin{cases} = 2^{-d}, & i = j, \\ \geq 2^{-d} + 2^{-d-2}, & i \neq j. \end{cases}$$

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

B Additional data analysis

B.1 Brain connectivity analysis

Figure S1 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 S1 and Table 2 (in the main paper), demonstrating the stability over CPT. One possible reason is that that our estimate is guaranteed to be in $[0, 1]$ (for binary tensor problem where $\mathcal{Y} \in \{0, 1\}^{d_1 \times \cdots \times d_K}$) whereas CPT estimation may fall outside the valid range $[0, 1]$.



Supplementary Figure S1: 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.

We next investigate the pattern in the estimated signal tensor. Figure 5 of the main paper shows the identified top edges associated with IQ scores. Specifically, we first obtain a denoised tensor $\hat{\Theta} \in \mathbb{R}^{68 \times 68 \times 114}$ using our method with $r = 10$ and $H = 20$. Then, we perform a regression analysis of $\hat{\Theta}(i, j, :) \in \mathbb{R}^{144}$ against the normalized IQ score across the 144 individuals. The regression model is repeated for each edge $(i, j) \in [68] \times [68]$. We find that top edges represent the interhemispheric connections in the frontal lobes. The result is consistent with the role of interhemispheric connectivity in human intelligence.

B.2 NIPS data analysis

In the main paper we have summarized the MAE in cross-validation experiments for $r = 6, 9, 12$. Here we provide additional results for a wider range $r = 3, 6, \ldots, 15$. 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 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			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. The strictly monotonicity of g implies that the inverse function $g^{-1}\colon \mathbb{R} \to \mathbb{R}$ is well-defined. When g is strictly increasing, the mapping $x \mapsto g(x)$ is sign preserving. Specifically, if $x \geq 0$, then $g(x) \geq g(0) = 0$. Conversely, if $g(x) \geq 0 = g(0)$, then applying g^{-1} to both sides gives $x \geq 0$. When g is strictly decreasing, the mapping $x \mapsto g(x)$ is sign reversing. Specifically, if $x \geq 0$, then $g(x) \leq g(0) = 0$. Conversely, if $g(x) \geq 0 = g(0)$, then applying g^{-1} to both sides gives $x \leq 0$. Therefore, $\Theta \simeq g(\Theta)$, or $\Theta \simeq -g(\Theta)$. Since constant multiplication does not change the tensor rank, we have $\mathrm{srank}(\Theta) = \mathrm{srank}(g(\Theta)) \leq \mathrm{rank}(g(\Theta))$. □

Proof of Proposition 2. See Section A.2 for constructive examples. □

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

$$\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta}) = \mathbb{E}_{\omega \sim \Pi} \underbrace{\mathbb{E}_{\mathcal{Y}(\omega)} \left\{ |\mathcal{Y}(\omega) - \pi| \left[\left| \mathcal{Z}(\omega) - \mathrm{sgn}(\bar{\mathcal{Y}}(\omega)) \right| - \left| \bar{\Theta}(\omega) - \mathrm{sgn}(\bar{\mathcal{Y}}(\omega)) \right| \right] \right\}}_{\stackrel{\mathrm{def}}{=} I(\omega)}. \quad (9)$$

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

$$\begin{aligned} I(\omega) &= \mathbb{E}_y \left[(y - \pi)(\bar{\theta} - z) \mathbb{1}(y \geq \pi) + (\pi - y)(z - \bar{\theta}) \mathbb{1}(y < \pi) \right] \\ &= \mathbb{E}_y \left[(\bar{\theta} - z)(y - \pi) \right] \\ &= \left[\mathrm{sgn}(\theta - \pi) - z \right] (\theta - \pi) \\ &= |\mathrm{sgn}(\theta - \pi) - z| |\theta - \pi| \geq 0, \end{aligned} \quad (10)$$

where the third line uses the fact $\mathbb{E}y = \theta$ and $\bar{\theta} = \mathrm{sgn}(\theta - \pi)$, and the last line uses the assumption $z \in \{-1, 1\}$. The equality (10) is attained when $z = \mathrm{sgn}(\theta - \pi)$ or $\theta = \pi$. Combining (10) with (9), we conclude that, for all $\mathcal{Z} \in \{-1, 1\}^{d_1 \times \cdots \times d_K}$,

$$\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta}) = \mathbb{E}_{\omega \sim \Pi} |\mathrm{sgn}(\Theta(\omega) - \pi) - \mathcal{Z}(\omega)| |\Theta(\omega) - \pi| \geq 0, \quad (11)$$

In particular, setting $\mathcal{Z} = \bar{\Theta} = \mathrm{sgn}(\Theta - \pi)$ in (11) yields the minimum. Therefore,

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

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

C.2 Proof of Theorem 1

Proof of Theorem 1. Fix $\pi \in [-1, 1]$. Based on (11) in Proposition 3 we have

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) = \mathbb{E}\left[|\text{sgn}\mathcal{Z} - \text{sgn}\bar{\Theta}||\bar{\Theta}|\right]. \tag{12}$$

The Assumption 1 states that

$$\mathbb{P}\left(|\bar{\Theta}| \le t\right) \le ct^{\alpha}, \quad \text{for all } 0 \le t < \rho(\pi, \mathcal{N}). \tag{13}$$

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

$$\begin{aligned}
\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) &\ge t\mathbb{E}\left(|\text{sgn}\mathcal{Z} - \text{sgn}\hat{\Theta}|\mathbb{1}\{|\hat{\Theta}| > t\}\right) \\
&\ge 2t\mathbb{P}\left(\text{sgn}\mathcal{Z} \ne \text{sgn}\bar{\Theta} \text{ and } |\bar{\Theta}| > t\right) \\
&\ge 2t\Big\{\mathbb{P}\left(\text{sgn}\mathcal{Z} \ne \text{sgn}\bar{\Theta}\right) - \mathbb{P}\left(|\bar{\Theta}| \le t\right)\Big\} \\
&\ge t\Big\{\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) - 2ct^{\alpha}\Big\},
\end{aligned} \tag{14}$$

  where the last line follows from the definition of MAE and (13). We maximize the lower bound (14) with respect to t, and obtain the optimal t_{opt},

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

  The corresponding lower bound of the inequality (14) becomes

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

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

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

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

- Case 2: $\alpha = \infty$. The inequality (14) now becomes

$$\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta}) \ge t\text{MAE}(\text{sgn}\bar{\Theta}, \text{sgn}\mathcal{Z}), \quad \text{for all } 0 \le t < \rho(\pi, \mathcal{N}). \tag{15}$$

  The conclusion follows by taking $t = \frac{\rho(\pi, \mathcal{N})}{2}$ in the inequality (15).

□

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

$$\text{MAE}(\text{sgn}\mathcal{Z}, \text{sgn}\bar{\Theta}) \lesssim \left[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})\right]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})}\left[\text{Risk}(\mathcal{Z}) - \text{Risk}(\bar{\Theta})\right], \tag{16}$$

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

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

$$\begin{aligned} L(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) &= \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \underbrace{|\bar{\mathcal{Y}}(\omega)|}_{\textit{weight}} \times \underbrace{|\mathrm{sgn}\mathcal{Z}(\omega) - \mathrm{sgn}\bar{\mathcal{Y}}(\omega)|}_{\textit{classification loss}} \\ &= \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}), \end{aligned} \tag{17}$$

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

$$\mathrm{Var}[\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)] \lesssim [\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})]^{\frac{\alpha}{1+\alpha}} + \frac{1}{\rho(\pi, \mathcal{N})}[\mathrm{Risk}(\mathcal{Z}) - \mathrm{Risk}(\bar{\Theta})], \tag{18}$$

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

Proof of Lemma C.1. We expand the variance by

$$\begin{aligned} \mathrm{Var}[\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)] &\lesssim \mathbb{E}|\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)|^2 \\ &\lesssim \mathbb{E}|\ell_\omega(\mathcal{Z}, \bar{\mathcal{Y}}_\Omega) - \ell_\omega(\bar{\Theta}, \bar{\mathcal{Y}}_\Omega)| \\ &\leq \mathbb{E}|\mathrm{sgn}\mathcal{Z} - \mathrm{sgn}\bar{\Theta}| = \mathrm{MAE}(\mathrm{sgn}\mathcal{Z}, \mathrm{sgn}\bar{\Theta}), \end{aligned} \tag{19}$$

where the second line comes from the boundedness of classification loss $L(\cdot, \cdot)$, and the third line comes from the inequality $||a - b| - |c - b|| \leq |a - b|$ for $a, b, c \in \{-1, 1\}$, together with the boundedness of classification weight $|\bar{\mathcal{Y}}(\omega)|$. Here we have absorbed the constant multipliers in $\lesssim$. The conclusion (18) then directly follows by applying Remark C.1 to (19). □

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

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

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

We next apply the empirical process theory to bound $\hat{\mathcal{Z}}$. To facilitate the analysis, we view the data $\bar{\mathcal{Y}}_\Omega = \{\bar{\mathcal{Y}}(\omega) : \omega \in \Omega\}$ 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 (17) becomes

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

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

$$L(\mathcal{Z},\bar{\mathcal{Y}}_\Omega)-L(\bar{\Theta},\bar{\mathcal{Y}}_\Omega)=\frac{1}{n}\sum_{i=1}^{n}\underbrace{\big[\ell_i(\mathcal{Z},\bar{\mathcal{Y}})-\ell_i(\bar{\Theta},\bar{\mathcal{Y}})\big]}_{\stackrel{\text{def}}{=}\Delta_i(f_{\mathcal{Z}},\bar{\mathcal{Y}})}, \tag{20}$$

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

Our remaining proof adopts the techniques of Wang et al. (2008, Theorem 3) to bound (20) over the function family $f_{\mathcal{Z}}\in\mathcal{F}_{\mathcal{T}}$. We summarize only the key difference here but refer to Wang et al. (2008) for complete proof. Based on Lemma C.1, the (α,π)-smoothness of Θ implies

$$\operatorname{Var}\Delta_i(f_{\mathcal{Z}},\bar{\mathcal{Y}})\lesssim\big[\mathbb{E}\Delta_i(f_{\mathcal{Z}},\bar{\mathcal{Y}})\big]^{\frac{\alpha}{1+\alpha}}+\frac{1}{\rho}\mathbb{E}\Delta_i(f_{\mathcal{Z}},\bar{\mathcal{Y}}),\quad\text{for all } f_{\mathcal{Z}}\in\mathcal{F}_{\mathcal{T}}. \tag{21}$$

Applying local iterative techniques in Wang et al. (2008, Theorem 3) to the empirical process (20) with the variance-to-mean relationship (21) gives that

$$\mathbb{P}\Big(\operatorname{Risk}(\hat{\mathcal{Z}})-\operatorname{Risk}(\bar{\Theta})\ge L_n\Big)\lesssim\exp(-nL_n), \tag{22}$$

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

$$\frac{1}{L_n}\int_{L_n}^{\sqrt{L_n^{\alpha/(\alpha+1)}+\frac{L_n}{\rho}}}\sqrt{\mathcal{H}_{[\,]}(\varepsilon,\mathcal{F}_{\mathcal{T}},\|\cdot\|_2)}d\varepsilon\le C\sqrt{n}, \tag{23}$$

for some constant $C>0$. In particular, the smallest L_n satisfying (23) 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 (23). Based on Lemma C.2, the inequality (23) is satisfied with

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

Therefore, by (22), with very high probability.

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

Inserting the above bound into (16) gives

$$\begin{aligned}\operatorname{MAE}(\operatorname{sgn}\hat{\mathcal{Z}},\operatorname{sgn}\bar{\Theta})&\lesssim[\operatorname{Risk}(\hat{\mathcal{Z}})-\operatorname{Risk}(\bar{\Theta})]^{\alpha/(\alpha+1)}+\frac{1}{\rho}[\operatorname{Risk}(\hat{\mathcal{Z}})-\operatorname{Risk}(\bar{\Theta})]\\&\lesssim t_n^{\alpha/(\alpha+2)}+\frac{1}{\rho^{\alpha/\alpha+1}}t_n^{\alpha/(\alpha+1)}+\frac{1}{\rho}t_n^{(\alpha+1)/(\alpha+2)}+\frac{1}{\rho^2}t_n\\&\le 4t_n^{\alpha/(\alpha+2)}+\frac{4}{\rho^2}t_n,\end{aligned} \tag{24}$$

where the last line follows from the fact that $a(b^2+b^{(\alpha+2)/(\alpha+1)}+b+1)\le 4a(b^2+1)$ with $a=\frac{t_n}{\rho^2}$ and $b=\rho t_n^{-1/(\alpha+2)}$. We plug t_n into (24) and absorb the term $K\log K$ into the constant. The conclusion is then proved. □

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

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

and

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

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

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

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

Then, the following inequality is satisfied.

$$\frac{1}{L_n}\int_{L_n}^{\sqrt{L_n^{\alpha/(\alpha+1)} + \frac{L_n}{\rho(\pi,\mathcal{N})}}} \sqrt{\mathcal{H}_{[\,]}(\varepsilon, \mathcal{F}_\mathcal{T}, \|\cdot\|_2)}\, d\varepsilon \le C n^{1/2}, \tag{25}$$

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

Proof of Lemma C.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 (25) gives

$$g(L) = \frac{1}{L}\int_L^{\sqrt{L^{\alpha/(\alpha+1)} + \rho^{-1}L}} \sqrt{d_{\max} r K \log\left(\frac{K}{\varepsilon}\right)}\, d\varepsilon. \tag{26}$$

By the monotonicity of the integrand in (26), we bound $g(L)$ by

$$\begin{aligned} g(L) &\le \frac{\sqrt{d_{\max} r K}}{L}\int_L^{\sqrt{L^{\alpha/(\alpha+1)} + \rho^{-1}L}} \sqrt{\log\left(\frac{K}{L}\right)}\, d\varepsilon \\ &\le \sqrt{d_{\max} r K (\log K - \log L)}\left(\frac{L^{\alpha/(2\alpha+2)} + \sqrt{\rho^{-1}L}}{L} - 1\right) \\ &\le \sqrt{d_{\max} r K \log K}\left(\frac{1}{L^{(\alpha+2)/(2\alpha+2)}} + \frac{1}{\sqrt{\rho L}}\right), \end{aligned} \tag{27}$$

where the second line follows from $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b > 0$. It remains to verify that $g(L_n) \le Cn^{1/2}$ for L_n specified in (25). Plugging L_n into the last line of (27) gives

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

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

$$\begin{aligned} \mathrm{MAE}(\hat{\Theta}, \Theta) &= \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi \in \Pi} \mathrm{sgn} \hat{Z}_\pi - \Theta \right| \\ &\le \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi \in \Pi} \Big(\mathrm{sgn} \hat{Z}_\pi - \mathrm{sgn}(\Theta - \pi) \Big) \right| + \mathbb{E} \left| \frac{1}{2H+1} \sum_{\pi \in \Pi} \mathrm{sgn}(\Theta - \pi) - \Theta \right| \\ &\le \frac{1}{2H+1} \sum_{\pi \in \Pi} \mathrm{MAE}(\mathrm{sgn} \hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)) + \frac{1}{H}, \end{aligned} \tag{28}$$

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

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

Write $n = |\Omega|$. Now it suffices to bound the first term in (28). We prove that

$$\frac{1}{2H+1} \sum_{\pi \in \Pi} \mathrm{MAE}(\mathrm{sgn} \hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)) \lesssim t_n^{\alpha/(\alpha+2)} + \frac{1}{H} + H t_n, \quad \text{with } t_n = \frac{d_{\max} r K \log K}{n}. \tag{29}$$

Theorem 2 implies that the sign estimation accuracy depends on the closeness of $\pi \in \mathcal{H}$ to the mass points in $\mathcal{H}$. Therefore, we partition the level set $\pi \in \mathcal{H}$ based on their closeness to $\mathcal{H}$. Specifically, let $\mathcal{N}_H \overset{\text{def}}{=} \bigcup_{\pi' \in \mathcal{N}} \left(\pi' - \frac{1}{H}, \pi' + \frac{1}{H} \right)$ denote the set of levels at least $\frac{1}{H}$-close to the mass points. We expand (29) by

$$\begin{aligned} &\frac{1}{2H+1} \sum_{\pi \in \Pi} \mathrm{MAE}(\mathrm{sgn} \hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)) \\ =&\frac{1}{2H+1} \sum_{\pi \in \Pi \cap \mathcal{N}_H} \mathrm{MAE}(\mathrm{sgn} \hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)) + \frac{1}{2H+1} \sum_{\pi \in \Pi \cap \mathcal{N}_H^c} \mathrm{MAE}(\mathrm{sgn} \hat{Z}_\pi, \mathrm{sgn}(\Theta - \pi)). \end{aligned} \tag{30}$$

By assumption, the first term involves only finite number of summands and thus can be bounded by $4C/(2H+1)$ where $C > 0$ is a constant such that $|\mathcal{N}| \le C$. We bound the second term using

the explicit forms of $\rho(\pi, \mathcal{N})$ in the sequence $\pi \in \Pi \cap \mathcal{N}_H^c$. Based on Theorem 2,

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

where the last inequality follows from the Lemma C.3. Combining the bounds for the two terms in (30) completes the proof for conclusion (29). Finally, plugging (29) into (28) yields

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

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

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

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

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

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

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

References

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 22*, 580–615.

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