SUPPLEMENTARY MATERIALS

Appendix: The appendix includes the proofs and extra simulation results.

Software and data: R-package contains code to perform the methods described in the article. The package also contains all datasets used as examples in the article.

A Extra numerical results

A.1 Results for Models 2 and 4 in Table 1

We first present simulation results for Models 2 and 4 omitted in Section 6. Figure S2 compares the estimation performance among the **Borda count, LSE**, and **Spectral methods**. We find that our Borda count algorithm outperforms others in both models. The first two columns in Figure S3 show the impact of the number of blocks k and degree of polynomial ℓ for the approximation with fixed dimension $d = 100$. Similar to results for Models 1, 3 and 5 in the main paper, we find the optimal k balances the trade-off between approximation error and signal tensor estimation error within each block. The last two columns compare our **Borda count** with other alternative methods. We find that our method still outperforms **LSE** and **Spectral** in all scenarios under Models 2 and 4.

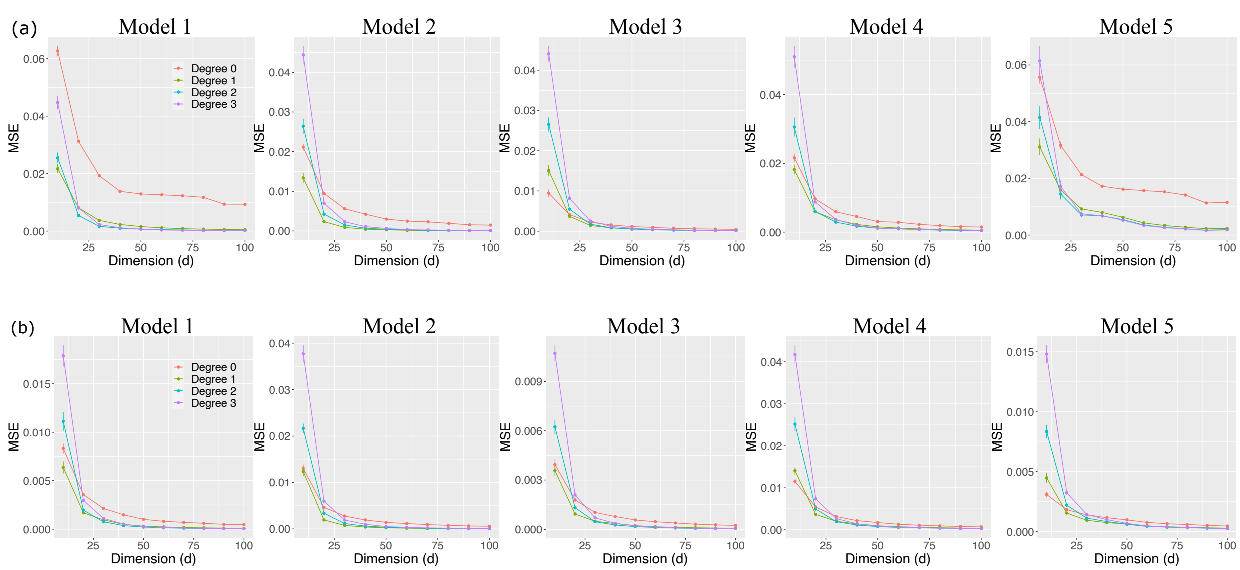

Figure S1: MSE versus the tensor dimension based on different polynomial approximations. Columns 1-5 consider the Models 1-5 in Table 2 respectively. Panel (a) is for continuous tensors, whereas (b) is for the binary tensors.

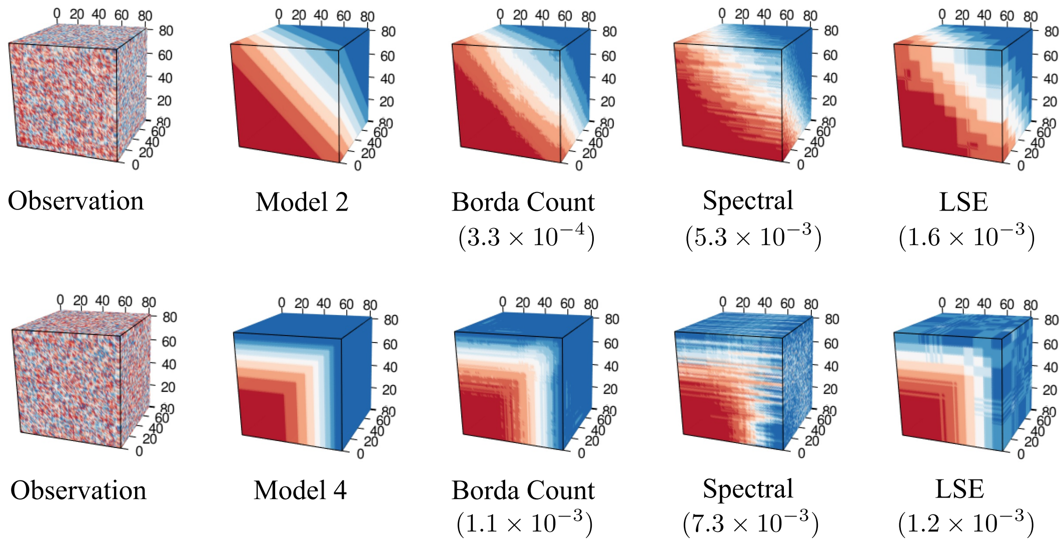

Figure S2: Performance comparison between different methods. The observed data tensors, true signal tensors, and estimated signal tensors are plotted for Models 2 and 4 in Table 2 with fixed dimension $d = 80$. Numbers in parenthesis indicate the mean squared error.

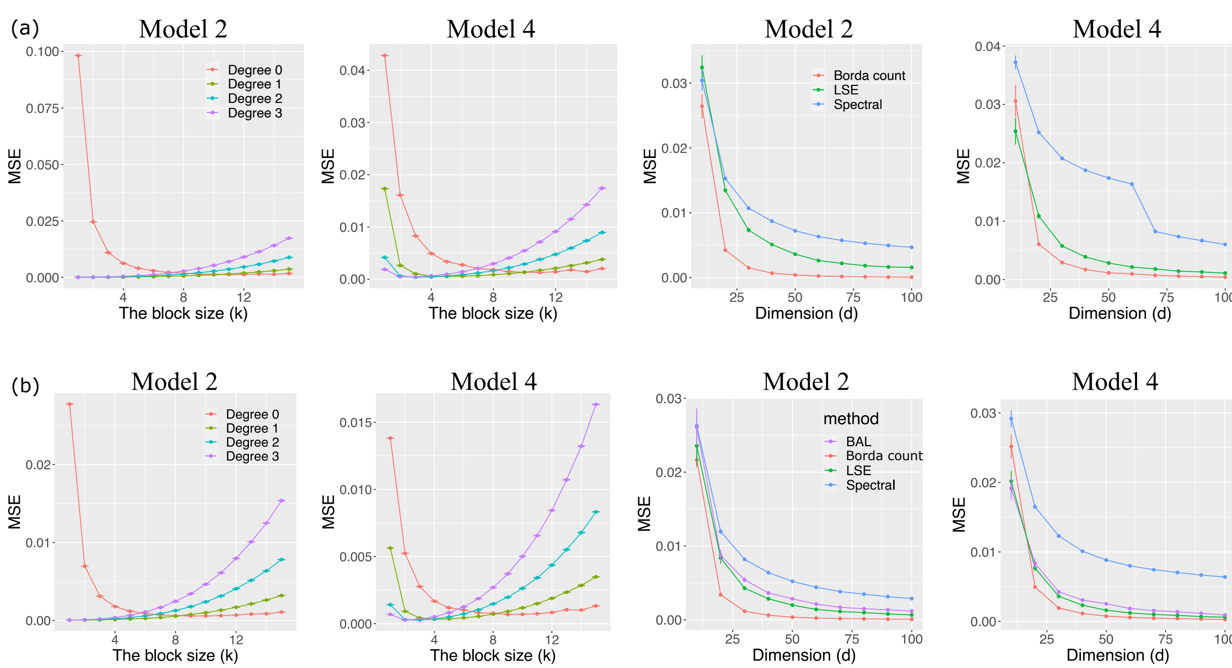

Figure S3: Simulation results for Models 2 and 4 in Table 2. Columns 1-2 plots MSE versus the number of blocks for different polynomial approximation, while Columns 3-4 shows the MSE versus the tensor dimension according to estimation methods. Panel (a) is for continuous tensors, whereas (b) is for the binary tensors.

A.2 Investigation of non-symmetric tensors

Here we describe the simulation set-up for non-symmetric tensors. We simulate order-3 tensors based on the non-symmetric functions in Table S1.

Model ID	$f(x,y,z)$
1	$xy+z$
2	x^2+y+yz^2
3	$x(1+\exp(-3(x^2+y^2+z^2)))^{-1}$
4	$\log(1+\max(x,y,z)+x^2+yz)$
5	$\exp\left(-x-\sqrt{y}-z^3\right)$

Table S1: List of non-symmetric smooth functions in simulation.

We fix the tensor dimension $30\times 40\times 50$ and assume that the noise tensors are from Gaussian distribution. Similar to other simulations, we evaluate the accuracy of the estimation by MSE and report the summary statistics across $n_{\text{sim}}=20$ replicates. The hyperparameters are chosen via cross-validation that give the best accuracy for each method. Table S2 summarizes the choice of hyperparameters.

Method	Model 1	Model 2	Model 3	Model 4	Model 5
Borda count	(2,1,2)	(1,2,2)	(1,3,3)	(2,1,2)	(1,4,4)
LSE	(6,2,3)	(8,5,8)	(6,9,6)	(9,5,6)	(7,9,3)
Spectral	(1,24)	(3,48)	(1,48)	(1,28)	(1,22)

Table S2: Hyperparameters for the methods under Models 1-5 in Table S1. For **Borda count** and **LSE** methods, the values in the table indicate the number of blocks. For **Spectral** method, the first value indicates the tensor unfolding mode, while the second one represents the singular value threshold.

A.3 Extra results on Chicago crime data analysis

We investigate the ten groups of crime types clustered by our method. Table S3 shows that the clustering captures the similar type of crimes. For example, group 2 consists of misdemeanors such as public indecency, non-criminal, and concealed carry license violation, while group 6 represents sex-related offenses such as prostitution, sex offense, and crime sexual assault.

GROUP	I	II	III
CRIME TYPE	RITUALISM, HUMAN TRAFFICKING, OTHER NARCOTIC VIOLATION	PUBLIC INDECENCY, NON-CRIMINAL, CONCEALED CARRY LICENSE VIOLATION	OBSCENITY, STALKING, INTIMIDATION
GROUP	IV	V	VI
CRIME TYPE	KIDNAPPING, GAMBLING, HOMICIDE	LIQUOR LAW VIOLATION, ARSON, INTERFERENCE WITH PUBLIC OFFICER	PROSTITUTION, SEX OFFENSE, CRIM SEXUAL ASSAULT
GROUP	VII	VIII	VIIII
CRIME TYPE	OTHER OFFENSE, CRIMINAL DAMAGE, BATTERY, THEFT, BURGLARY	CRIMINAL TRESPASS, ROBBERY, DECEPTIVE PRACTICE	NARCOTICS, ASSAULT, MOTOR VEHICLE THEFT
GROUP	X		
CRIME TYPE	PUBLIC PEACE VIOLATION, WEAPONS VIOLATION, OFFENSE INVOLVING CHILDREN		

Table S3: Groups of crime types learned based on the Borda count estimation.

B Proofs of main theorems

B.1 Proof of Lemma 1

Proof. Recall that we denote $\mathcal{E}_k$ as the m-way partition

$$\mathcal{E}_k = \{\bigtimes_{a=1}^m z^{-1}(j_a) \colon (j_1,\ldots,j_m)\in[k]^m\},$$

where $z\colon [d]\to[k]$ is the canonical clustering function such that $z(i)=\lceil ki/d\rceil$, for all $i\in[d]$, and we use the shorthand $\bigtimes_{a=1}^m$ to denote the Cartesian product of m sets. For a given partition $\bigtimes_{a=1}^m z^{-1}(j_a)\in\mathcal{E}_k$, fix any index $(i_1^0,\ldots,i_m^0)\in\bigtimes_{a=1}^m z^{-1}(j_a)$. Then, we have

$$\|(i_1,\ldots,i_m)-(i_1^0,\ldots,i_m^0)\|_\infty \leq \frac{d}{k}, \tag{18}$$

for all $(i_1,\ldots,i_m)\in\bigtimes_{a=1}^m z^{-1}(j_a)$. We define the block-wise degree-ℓ polynomial tensor $\mathcal{B}$ based on the partition $\mathcal{E}_k$ as

$$\mathcal{B}(i_1,\ldots,i_m)=\text{Poly}^{j_1,\ldots,j_m}_{\min(\lfloor\alpha\rfloor,\ell)}\left(\frac{i_1-i_1^0}{d},\ldots\frac{i_m-i_m^0}{d}\right),\quad \text{for all } (i_1,\ldots,i_m)\in\bigtimes_{a=1}^m z^{-1}(j_a),$$

where $\text{Poly}^{j_1,\ldots,j_m}_{\min(\lfloor\alpha\rfloor,\ell)}$ denotes a degree-ℓ polynomial function satisfying

$$\left|f\left(\frac{i_1}{d},\ldots,\frac{i_m}{d}\right)-\text{Poly}^{j_1,\ldots,j_m}_{\min(\lfloor\alpha\rfloor,\ell)}\left(\frac{i_1-i_1^0}{d},\ldots\frac{i_m-i_m^0}{d}\right)\right|\leq L\left\|\left(\frac{i_1-i_1^0}{d},\ldots,\frac{i_m-i_m^0}{d}\right)\right\|_\infty^{\min(\alpha,\ell+1)}, \tag{19}$$

for all $(i_1,\ldots,i_m)\in\bigtimes_{a=1}^m z^{-1}(j_a)$. Notice that we can always find such polynomial function by α-Hölder smoothness of the generative function f. Based on the construction of block-wise degree-ℓ polynomial tensor $\mathcal{B}$, we have

$$\begin{aligned}
&\frac{1}{d^m}\|\Theta-\mathcal{B}\|_F^2\\
&=\frac{1}{d^m}\sum_{(i_1,\ldots,i_m)\in[d]^m}|\Theta(i_1,\ldots,i_m)-\mathcal{B}(i_1,\ldots,i_m)|^2\\
&=\frac{1}{d^m}\sum_{(j_1,\ldots,j_m)\in[k]^m}\sum_{(i_1,\ldots,i_m)\in\times_{a=1}^m z^{-1}(j_a)}\left|f\left(\frac{i_1}{d},\ldots,\frac{i_m}{d}\right)-\text{Poly}^{j_1,\ldots,j_m}_{\min(\lfloor\alpha\rfloor,\ell)}\left(\frac{i_1-i_1^0}{d},\ldots,\frac{i_m-i_m^0}{d}\right)\right|^2\\
&\lesssim\frac{L^2}{d^m}\sum_{(j_1,\ldots,j_m)\in[k]^m}\sum_{(i_1,\ldots,i_m)\in\times_{a=1}^m z^{-1}(j_a)}\left\|\left(\frac{i_1-i_1^0}{d},\ldots,\frac{i_m-i_m^0}{d}\right)\right\|_\infty^{2\min(\alpha,\ell+1)}\\
&\leq\frac{L^2}{k^{2\min(\alpha,\ell+1)}},
\end{aligned}$$

where the first inequality uses (19) and the second inequality is from (18). □

B.2 Proof of Theorem 1

Proof. By Lemma 1, there exists a block-wise polynomial tensor $\mathcal{B}\in\mathscr{B}(k,\ell)$ such that

$$\|\mathcal{B}-\Theta\|_F^2\lesssim\frac{L^2d^m}{k^{2\min(\alpha,\ell)}}. \tag{20}$$

By the triangle inequality,

$$\|\hat{\Theta}^{\text{LSE}}\circ\hat{\pi}^{\text{LSE}}-\Theta\circ\pi\|_F^2\leq 2\|\hat{\Theta}^{\text{LSE}}\circ\hat{\pi}^{\text{LSE}}-\mathcal{B}\circ\pi\|_F^2+2\underbrace{\|\mathcal{B}\circ\pi-\Theta\circ\pi\|_F^2}_{\text{Lemma 1}}. \tag{21}$$

Therefore, it suffices to bound $\|\hat{\Theta}^{\text{LSE}}\circ\hat{\pi}^{\text{LSE}}-\mathcal{B}\circ\pi\|_F^2$. By the global optimality of least-square estimator, we have

$$\begin{aligned}
\|\hat{\Theta}^{\text{LSE}}\circ\hat{\pi}^{\text{LSE}}-\mathcal{B}\circ\pi\|_F&\leq\left\langle\frac{\hat{\Theta}^{\text{LSE}}\circ\hat{\pi}^{\text{LSE}}-\mathcal{B}\circ\pi}{\|\hat{\Theta}^{\text{LSE}}\circ\hat{\pi}^{\text{LSE}}-\mathcal{B}\circ\pi\|_F},\ \mathcal{E}+(\Theta\circ\pi-\mathcal{B}\circ\pi)\right\rangle\\
&\leq\sup_{\pi,\pi'\colon[d]\to[d]}\sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)}\left\langle\frac{\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi}{\|\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi\|_F},\mathcal{E}\right\rangle+\underbrace{\|\mathcal{B}\circ\pi-\Theta\circ\pi\|_F}_{\text{Lemma 1}}.
\end{aligned}$$

Now we bound inner product term. For fixed π,π', let $\boldsymbol{P}$ and $\boldsymbol{P}'$ be permutation matrices corresponding to permutations π and π' respectively. We express vectorized block-wise degree-ℓ polynomial tensors, vec($\mathcal{B}$) and vec($\mathcal{B}'$), by discrete polynomial functions. Specifically, denote vec$(\mathcal{B})=\boldsymbol{X}\boldsymbol{\beta}$ and vec$(\mathcal{B}')=\boldsymbol{X}\boldsymbol{\beta}'$, where $\boldsymbol{X}\in\mathbb{R}^{d^m\times k^m(k+m)^\ell}$ is a design matrix consisting of m-multivaraite degree-ℓ polynomial basis over grid design $(1/d,\ldots,d/d)$, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'\in\mathbb{R}^{k^m(k+m)^\ell}$ are corresponding coefficient vectors. Notice that the number of coefficients for m-multivariate polynomial of degree-ℓ is $\binom{\ell+m}{\ell}$. We choose to use $(k+m)^\ell$ coefficients for each block for notational simplicity. Therefore, we rewrite the inner product

$$\begin{aligned}\left\langle \frac{\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi}{\|\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi\|_F},\mathcal{E}\right\rangle &= \left\langle \frac{(\boldsymbol{P}')^{\otimes m}\mathrm{vec}(\mathcal{B}')-(\boldsymbol{P})^{\otimes m}\mathrm{vec}(\mathcal{B})}{\|(\boldsymbol{P}')^{\otimes m}\mathrm{vec}(\mathcal{B}')-(\boldsymbol{P})^{\otimes m}\mathrm{vec}(\mathcal{B})\|_F},\mathcal{E}\right\rangle \\ &= \left\langle \frac{(\boldsymbol{P}')^{\otimes m}\boldsymbol{X}\boldsymbol{\beta}'-(\boldsymbol{P})^{\otimes m}\boldsymbol{X}\boldsymbol{\beta}}{\|(\boldsymbol{P}')^{\otimes m}\boldsymbol{X}\boldsymbol{\beta}'-(\boldsymbol{P})^{\otimes m}\boldsymbol{X}\boldsymbol{\beta}\|_F},\mathcal{E}\right\rangle \\ &= \left\langle \frac{\boldsymbol{A}\boldsymbol{c}}{\|\boldsymbol{A}\boldsymbol{c}\|_F},\mathcal{E}\right\rangle,\end{aligned}$$

where we define $\boldsymbol{A} := \begin{pmatrix}\boldsymbol{P}' & -\boldsymbol{P}\end{pmatrix}\begin{pmatrix}\boldsymbol{X} & 0\\ 0 & \boldsymbol{X}\end{pmatrix} \in \mathbb{R}^{d^m\times 2k^m(k+m)^\ell}$ and $\boldsymbol{c} := \begin{pmatrix}\boldsymbol{\beta}'\\ \boldsymbol{\beta}\end{pmatrix} \in \mathbb{R}^{2k^m(k+m)^\ell}$. By Lemma 5, we have

$$\sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)}\left\langle \frac{\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi}{\|\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi\|_F},\mathcal{E}\right\rangle \leq \sup_{\boldsymbol{c}\in\mathbb{R}^{2k^m(\ell+m)^\ell}}\left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2},\ e\right\rangle, \tag{22}$$

where $e\in\mathbb{R}^{2k^m(k+m)^\ell}$ is a vector consisting of i.i.d. sub-Gaussian entries with variance proxy σ^2. By the union bound of Gaussian maxima over countable set $\{\pi,\pi'\colon [d]\to[d]\}$, we obtain

$$\begin{aligned}&\mathbb{P}\left(\sup_{\pi,\pi'\colon[d]\to[d]}\sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)}\left\langle \frac{\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi}{\|\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi\|_F},\mathcal{E}\right\rangle \geq t\right) \\ &\leq \sum_{\pi,\pi'\in[d]^d}\mathbb{P}\left(\sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)}\left\langle \frac{\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi}{\|\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi\|_F},\mathcal{E}\right\rangle \geq t\right) \\ &\leq d^d\mathbb{P}\left(\sup_{\boldsymbol{c}\in\mathbb{R}^{2k^m(\ell+m)^\ell}}\left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2},\ e\right\rangle \geq t\right) \\ &\leq \exp\left(-\frac{t^2}{8\sigma^2}+k^m(\ell+m)^\ell\log 6+d\log d\right),\end{aligned} \tag{23}$$

where the second inequality is from (22) and the last inequality is from Lemma 6. Setting $t=C\sigma\sqrt{k^m(\ell+m)^\ell+d\log d}$ in (24) for sufficiently large $C>0$ gives

$$\sup_{\pi,\pi'\colon[d]\to[d]}\sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)}\left\langle \frac{\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi}{\|\mathcal{B}'\circ\pi'-\mathcal{B}\circ\pi\|_F},\mathcal{E}\right\rangle \lesssim \sigma\sqrt{k^m(\ell+m)^\ell+d\log d}, \tag{24}$$

with high probability.

Combining the inequalities (20), (21) and (24) yields the desired conclusion

$$\|\hat{\Theta}^{\mathrm{LSE}}\circ\hat{\pi}^{\mathrm{LSE}}-\Theta\circ\pi\|_F^2 \lesssim \sigma^2\left(k^m(\ell+m)^\ell+d\log d\right)+\frac{L^2d^m}{k^{2\min(\alpha,\ell)}}. \tag{25}$$

Finally, optimizing (25) with respect to (k,l) gives that

$$(25)\ \lesssim \begin{cases} L^2\left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}}d^{-\frac{2m\alpha}{m+2\alpha}}, & \text{when } \alpha < m(m-1)/2,\\ \sigma^2 d^{-(m-1)}\log d, & \text{when } \alpha\geq m(m-1)/2,\end{cases}$$

under the choice

$$\ell^* = \min(\lfloor\alpha\rfloor, (m-2)(m+1)/2), \quad k^* = \left\lceil \left(d^m L^2/\sigma^2\right)^{\frac{1}{m+2\min(\alpha,\ell^*+1)}} \right\rceil.$$

□

B.3 Proof of Theorem 2

Proof. By the definition of the tensor space, we seek the minimax rate ε^2 in the following expression

$$\inf_{(\hat{\Theta},\hat{\pi})} \sup_{\Theta\in\mathcal{P}(\alpha,L)} \sup_{\pi\in\Pi(d,d)} \mathbb{P}\left(\frac{1}{d^m}\|\Theta\circ\pi - \hat{\Theta}\circ\hat{\pi}\|_F^2 \geq \varepsilon^2\right).$$

On one hand, if we fix a permutation $\pi \in \Pi(d,d)$, the problem can be viewed as a classical m-dimensional α-smooth nonparametric regression with d^m sample points. The minimax lower bound is known to be $\varepsilon^2 = L^2\left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}} d^{-\frac{2m\alpha}{m+2\alpha}}$. On the other hand, if we fix $\Theta \in \mathcal{P}(\alpha, L)$, the problem become a new type of convergence rate due to the unknown permutation. We refer to the resulting error as the permutation rate, and we will prove that $\varepsilon^2 = \sigma^2 d^{-(m-1)}\log d$. Since our target is the sum of the two rates, it suffice to prove the two different rates separately. In the following arguments, we will proceed by this strategy.

Nonparametric rate. The nonparametric rate for α-smooth function is readily available in the literature; see Györfi et al. [11, Section 3.2] and Stone [25, Section 2]. We state the results here for self-completeness.

Lemma 3 (Minimax rate for α-smooth function estimation)**.** *Consider a sample of N data points, $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_N, Y_N)$, where $\boldsymbol{x}_n = (\frac{i_1}{d}, \ldots, \frac{i_m}{d}) \in [0,1]^m$ is the m-dimensional predictor and $Y_n \in \mathbb{R}$ is the scalar response. Consider the observation model*

$$Y_n = f(\boldsymbol{x}_n) + \varepsilon_n, \quad \textit{with } \varepsilon_n \sim \textit{i.i.d. } N(0,1), \quad \textit{for all } n \in [N].$$

Assume f is in the α-Holder smooth function class, denoted by $\mathcal{H}(\alpha, L)$. Then,

$$\inf_{\hat{f}} \sup_{f\in\mathcal{H}(\alpha,L)} \mathbb{P}\left(\|f-\hat{f}\|_2 \geq \sigma^{\frac{4\alpha}{m+2\alpha}} L^{\frac{2m}{m+2\alpha}} N^{-\frac{2\alpha}{m+2\alpha}}\right) \geq 0.9.$$

Our desired nonparametric rate readily follows from Lemma 3 by taking sample size $N = d^m$ and function norm $\|f - \hat{f}\|_2 = \frac{1}{d^m}\|\Theta - \hat{\Theta}\|_F^2$. In summary, for a given permutation $\pi \in \Pi(d,d)$, we have

$$\inf_{\hat{\Theta}} \sup_{\Theta\in\mathcal{P}(\alpha,L)} \mathbb{P}\left(\frac{1}{d^m}\|\hat{\Theta}\circ\pi - \Theta\circ\pi\|_F^2 \geq L^2\left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}} d^{-\frac{2m\alpha}{m+2\alpha}}\right) \geq 0.9. \tag{26}$$

Permutation rate. Since nonparametric rate dominates permutation rate when $\alpha \leq 1$, it is sufficient to prove the permutation rate lower bound for $\alpha \geq 1$. We first show the minimax permutation rate for k-block degree-0 tensor family $\mathcal{B}(k,0)$, and then construct a smooth $f \in \mathcal{H}(\alpha, L)$ to mimic the constant block tensors.

Let $\Pi(d,k)$ denote the collection of all possible onto mappings from $[d]$ to $[k]$. Lemma 4 shows the permutation rate over k-block degree-0 tensor family $\mathcal{B}(k,0)$ is $\sigma^2 d^{-(m-1)} \log k$.

Lemma 4 (Permutation error for tensor block model)**.** *Consider the problem of estimating d-dimensional, block-k signal tensors from sub-Gaussian tensor block models. For every given integer $k \in [d]$, there exists a core tensor $\mathcal{S} \in \mathbb{R}^{k\times\cdots\times k}$ satisfying*

$$\inf_{\hat{\Theta}} \sup_{z\in\Pi(d,k)} \mathbb{P}\left\{ \frac{1}{d^m} \sum_{(i_1,\ldots,i_m)\in[d]^m} \left[\hat{\Theta}(i_1,\ldots,i_m) - \mathcal{S}\left(z(i_1),\ldots,z(i_m)\right)\right]^2 \gtrsim \frac{\sigma^2 \log k}{d^{m-1}} \right\} \geq 0.9. \tag{27}$$

The proof of Lemma 4 is constructive and deferred to next subsection. We fix a core tensor $\mathcal{S} \in \mathbb{R}^{k\times\cdots\times k}$ satisfying (27), and use it to construct the smooth tensors.

Now we construct a function $f \in \mathcal{H}(\alpha, L)$ that mimics the core tensor $\mathcal{S}$ in block tensor family $\mathcal{B}(k,0)$. Define $k = d^\delta$ for some $\delta \in (0,1)$, which will be specified later. Consider a smooth function $K(x)$ that is infinitely differentiable,

$$K(x) = C_k \exp\left(-\frac{1}{1-64x^2}\right) \mathbb{1}\left\{|x| < \frac{1}{8}\right\},$$

where $C_k > 0$ satisfies $\int K(x)dx = 1$. Then, we define a smooth cutoff function as

$$\psi(x) = \int_{-3/8}^{3/8} K(x-y)dy.$$

The smooth cutoff function has support $[-1/2, 1/2]$ and takes value 1 on the interval $[-1/4, 1/4]$. For a given core tensor $\mathcal{S}$ satisfying Lemma 4, we define α-smooth function

$$f(x_1,\ldots,x_m) = \sum_{(a_1,\ldots,a_m)\in[k]^m} \left(\mathcal{S}(a_1,\ldots,a_m) - \frac{1}{2}\right) \prod \psi\left(kx_1 - a_1 + \frac{1}{2}\right) + \frac{1}{2}. \tag{28}$$

One can verify that $f \in \mathcal{H}(\alpha, L)$ as long as we choose sufficiently small δ depending on α and L. Notice that for any $(a_1,\ldots,a_m) \in [k]^m$,

$$f(x_1,\ldots,x_m) = \mathcal{S}(a_1,\ldots,a_m), \quad \text{if } (x_1,\ldots,x_m) \in \bigtimes_{i=1}^m \left[\frac{a_i - 3/4}{k}, \frac{a_i - 1/4}{k}\right].$$

From this observation, we define a sub-domain $I \subset [d]$ such that

$$I = \left(\bigcup_{a=1}^k \left[\frac{d(a-3/4)}{k}, \frac{d(a-1/4)}{k}\right]\right) \bigcap [d].$$

Then, $\{f(i_1/d,\ldots,i_m/d)\colon i_1,\ldots,i_m\in I\}$ forms the block structure with the core tensor $\mathcal{S}\in\mathbb{R}^{k\times\cdots k}$. Define a subset of permutations $\Pi'(d,d)=\{\pi\in\Pi(d,d)\colon \sigma(i)=i \text{ for } i\in[d]\setminus I\}\subset\Pi(d,d)$, which collects permutations on I while fixing indices on $[d]\setminus I$. Then we have

$$
\begin{aligned}
&\inf_{(\hat{\Theta},\hat{\pi})}\sup_{\pi\in\Pi(d,d)}\mathbb{P}\left(\frac{1}{d^m}\|\hat{\Theta}\circ\hat{\pi}-\Theta\circ\pi\|_F^2\geq\varepsilon^2\right)\\
\overset{(1)}{=}\ &\inf_{\hat{\Theta}}\sup_{\pi\in\Pi(d,d)}\mathbb{P}\left(\frac{1}{d^m}\|\hat{\Theta}-\Theta\circ\pi\|_F^2\geq\varepsilon^2\right)\\
\overset{(2)}{\geq}\ &\inf_{\hat{\Theta}}\sup_{\pi\in\Pi'(d,d)}\mathbb{P}\left(\frac{1}{d^m}\sum_{(i_1,\ldots,i_m)\in[d]^m}\left[\hat{\Theta}(i_1,\ldots,i_m)-f(\pi(i_1)/d,\ldots,\pi(i_m)/d)\right]^2\geq\varepsilon^2\right)\\
\geq\ &\inf_{\hat{\Theta}}\sup_{\pi\in\Pi'(d,d)}\mathbb{P}\left(\frac{1}{d^m}\sum_{(i_1,\ldots,i_m)\in I^m}\left[\hat{\Theta}(i_1,\ldots,i_m)-f(\pi(i_1)/d,\ldots,\pi(i_m)/d)\right]^2\geq\varepsilon^2\right),
\end{aligned}\tag{29}
$$

where (1) absorbs the estimate $\hat{\pi}$ into the estimate $\hat{\Theta}$, and (2) uses the constructed function (28) and the permutation collections $\Pi'(d,d)$. For any $\pi\in\Pi'(d,d)$, define clustering function $z\colon I\to[k]$ such that $z(i)=\lceil k\pi(i)/d\rceil$ for all $i\in I$. Then, we have

$$
f\left(\frac{\pi(i_1)}{d},\ldots,\frac{\pi(i_m)}{d}\right)=\mathcal{S}(z(i_1),\ldots,z(i_m)),\quad\text{for all } i_1,\ldots,i_m\in I.\tag{30}
$$

Finally, combining (29), (30), and Lemma 4 yields

$$
\inf_{(\hat{\Theta},\hat{\pi})}\sup_{\pi\in\Pi(d,d)}\mathbb{P}\left(\frac{1}{d^m}\|\hat{\Theta}\circ\hat{\pi}-\Theta\circ\pi\|_F^2\gtrsim\frac{\sigma^2\log d}{d^{m-1}}\right)\geq 0.9,\tag{31}
$$

where k is replaced by n^δ.

Combining two rates. Now, we combine (26) and (31) to get the desired lower bound. For any Θ generated as in (3) with $f\in\mathcal{H}(\alpha,L)$, by union bound, we have

$$
\begin{aligned}
&\mathbb{P}\left\{\frac{1}{d^m}\|\hat{\Theta}-\Theta\|_F^2\gtrsim L^2\left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}}d^{-\frac{2m\alpha}{m+2\alpha}}+\frac{\sigma^2\log d}{d^{m-1}}\right\}\\
\geq\ &\mathbb{P}\left\{\frac{1}{d^m}\|\hat{\Theta}-\Theta\|_F^2\gtrsim L^2\left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}}d^{-\frac{2m\alpha}{m+2\alpha}}\right\}+\mathbb{P}\left\{\frac{1}{d^m}\|\hat{\Theta}-\Theta\|_F^2\gtrsim\frac{\sigma^2\log d}{d^{m-1}}\right\}-1.
\end{aligned}
$$

Taking sup on both sides with the property

$$
\sup_{\substack{\Theta\in\mathcal{P}(\alpha,L)\\ \pi\in\Pi(d,d)}}(f(\pi)+g(\Theta))=\sup_{\pi\in\Pi(d,d)}f(\pi)+\sup_{\Theta\in\mathcal{P}(\alpha,L)}g(\Theta)
$$

yields the desired rate (11). □

B.4 Proof of Lemma 4

Proof. We provide the proof for $m = 3$ only. The extension to higher orders ($m \geq 4$) uses exactly the same techniques and thus is omitted. Let us pick $\omega_1, \ldots, \omega_{k/3} \in \{0,1\}^{k^2/9}$ such that $\rho_H(\omega_p, \omega_q) \geq k^2/36$ for all $p \neq q \in [k/3]$. This selection is possible by lemma 7. Fixing such $\omega_1, \ldots, \omega_{k/3}$, we define a symmetric core tensor $\mathcal{S} \in \mathbb{R}^{k\times k\times k}$ for $p < q < r$,

$$\mathcal{S}(p,q,r) = \begin{cases} s_{p,q,r} & \text{if } p \in \{1,\ldots,k/3\}, q \in \{k/3+1,\ldots,2k/3\}, r \in \{2k/3+1,\ldots,k\}, \\ 0 & \text{Otherwise,} \end{cases}$$

where $\{s_{p,q,r} \colon p \in \{1,\ldots,k/3\}, q \in \{k/3+1,\ldots,2k/3\}, r \in \{2k/3+1,\ldots,k\}\}$ satisfies

$$\begin{aligned} \boldsymbol{s}(r) &:= \operatorname{vec}\left(\mathcal{S}\left(1 : \frac{k}{3},\ \frac{k}{3}+1 : \frac{2k}{3},\ r\right)\right) \qquad (32) \\ &= \sqrt{\frac{c\sigma^2 \log k}{d^2}}\omega_{r-2k/3} \qquad \text{for any } r \in \{2k/3+1,\ldots,k\}. \end{aligned}$$

The choice of constant $c > 0$ is deferred to a later part of the proof. Notice that for any $r_1, r_2 \in \{2k/3+1,\ldots,k\}$, we have

$$\|\boldsymbol{s}(r_1) - \boldsymbol{s}(r_2)\|_F^2 \geq \frac{c\sigma^2 k^2 \log k}{36 d^2}. \qquad (33)$$

Define a subset of permutation set $\Pi(d,k)$ by

$$\mathcal{Z} = \left\{ z \in \Pi(d,k) \colon |z^{-1}(p)| = \frac{d}{k} \text{ for } a \in [k], z^{-1}(a) = \left\{ \frac{(p-1)d}{k} + 1, \ldots, \frac{pd}{k} \text{ for } p \in [2k/3] \right\} \right\}.$$

Each $z \in \mathcal{Z}$ induces a block tensor in $\mathcal{B}(k,0)$. We consider the collection of block tensors induced by $\mathcal{Z}$; i.e.,

$$\mathcal{B}(\mathcal{Z}) = \{\Theta^z \in \mathbb{R}^{d\times d\times d} \colon \Theta^z(i,j,k) = \mathcal{S}(z(i), z(j), z(k)) \text{ for } z \in \mathcal{Z}\}.$$

To apply Proposition 1, we find upper bound $\sup_{\Theta,\Theta' \in \mathcal{B}(\mathcal{Z})} D(\mathbb{P}_\Theta | \mathbb{P}_{\Theta'})$ and lower bound $\log \mathcal{M}(\epsilon, \mathcal{B}(\mathcal{Z}), \rho)$, where ρ is defined by $\rho(\Theta, \Theta') = \frac{1}{n^3}\|\Theta - \Theta'\|_F^2$. For sub-Gaussian signal plus noise model, we have

$$D(\mathbb{P}_\Theta | \mathbb{P}_{\Theta'}) \leq \frac{1}{2\sigma^2}\|\Theta - \Theta'\|_F \leq \frac{1}{2\sigma^2} d^3 \frac{c\sigma^2 \log k}{d^2} = \frac{cd \log k}{2}, \qquad (34)$$

where the first inequality holds for any $\Theta, \Theta' \in \mathcal{B}(\mathcal{Z})$ by Gao et al. [10, Proposition 4.2]. Now we provide a lower bound of the packing number $\log \mathcal{M}(\epsilon, \mathcal{B}(\mathcal{Z}), \epsilon)$ with $\epsilon^2 \asymp \frac{\sigma^2 \log k}{d^2}$. From the construction of $\mathcal{S}$ in (32), we have one to one correspondence between $\mathcal{Z}$ and $\mathcal{B}(\mathcal{Z})$. Thus $\mathcal{M}(\epsilon, \mathcal{B}(\mathcal{Z}), \rho) = \mathcal{M}(\epsilon, \mathcal{Z}, \rho')$ for some metric ρ' on $\mathcal{Z}$ defined by $\rho'(z_1, z_2) = \rho(\Theta^{z_1}, \Theta^{z_2})$. Let P be the packing set in $\mathcal{Z}$ with the same cardinality of $\mathcal{M}(\epsilon, \mathcal{Z}, \rho')$. Given any $z \in \mathcal{Z}$, define its ϵ-neighbor by $\mathcal{N}(z,\epsilon)=\{z'\in\mathcal{Z}\colon \rho'(z,z')\leq\epsilon\}$. Then, we have $\cup_{z\in P}\mathcal{N}(z,\epsilon)=\mathcal{Z}$, because the cardinality of P is same as packing number $\mathcal{M}(\epsilon,\mathcal{Z},\rho')$. Therefore, we have

$$|\mathcal{Z}|\leq\sum_{z\in P}|\mathcal{N}(z,\epsilon)|\leq|P|\max_{z\in P}|\mathcal{N}(z,\epsilon)|. \tag{35}$$

It remains to find the upper bound of $\max_{z\in P}|\mathcal{N}(z,\epsilon)|$. For any $z_1,z_2\in\mathcal{Z}$, $z_1(i)=z_2(i)$ for $i\in[2d/3]$ and $|z_1^{-1}(p)|=d/k$ for all $p\in[k]$. Therefore,

$$\begin{aligned}
\rho'^2(z_1,z_2)&\geq\frac{1}{d^3}\sum_{1\leq i_1\leq d/3<i_2\leq 2d/3\leq i_3\leq d}\left(\mathcal{S}\left(z_1(i_1),z_1(i_2),z_1(i_3)\right)-\mathcal{S}\left(z_2(i_1),z_2(i_2),z_2(i_3)\right)\right)^2\\
&=\frac{1}{d^3}\sum_{2n/3<i_3\leq n}\sum_{1\leq p\leq k/3<q\leq 2k/3}\sum_{i_1\in z_1^{-1}(p),i_2\in z_1^{-1}(q)}\left(\mathcal{S}\left(p,q,z_1(i_3)\right)-\mathcal{S}\left(p,q,z_2(i_3)\right)\right)^2\\
&=\frac{1}{d^3}\sum_{2n/3<i_3\leq n}\sum_{1\leq p\leq k/3<q\leq 2k/3}\left(\frac{d}{k}\right)^2\left(\mathcal{S}\left(p,q,z_1(i_3)\right)-\mathcal{S}\left(p,q,z_2(i_3)\right)\right)^2\\
&=\frac{1}{d^3}\sum_{2n/3<i_3\leq n}\left(\frac{d}{k}\right)^2\|\boldsymbol{s}(z_1(i_3)-\boldsymbol{s}(z_2(i_3))\|_F^2\\
&\geq\frac{c\sigma^2\log k}{36d^3}|\{j\colon z_1(j)\neq z_2(j)\}|,
\end{aligned}$$

where the last inequality is from (33). Hence with the choice of $\epsilon^2=\frac{c\sigma^2\log k}{288d^2}$, we have $|\{j\colon z(j)\neq z'(j)\}|\leq d/8$ for any $z'\in\mathcal{N}(z,\epsilon)$. This implies

$$|\mathcal{N}(z,\epsilon)|\leq\binom{d}{d/8}k^{d/8}\leq(8e)^{d/8}k^{d/8}\leq\exp\left(\frac{1}{5}d\log k\right), \tag{36}$$

for sufficiently large k. Now we find the lower bound of $|\mathcal{Z}|$ based on Stirling's formula,

$$|\mathcal{Z}|=\frac{(d/3)!}{[(d/k)!]^{k/3}}=\exp\left(\frac{1}{3}d\log k+o(d\log k)\right)\geq\exp\left(\frac{1}{4}d\log k\right). \tag{37}$$

Plugging (36) and (37) into (35) yields

$$\mathcal{M}(\epsilon,\mathcal{B}(\mathcal{Z}),\rho)=|P|\geq\frac{\max_{z\in P}\mathcal{N}(z,\epsilon)}{|\mathcal{Z}|}\geq\exp\left(\frac{1}{20}d\log k\right). \tag{38}$$

Finally, applying Proposition 1 based on (34) and (38) gives

$$\inf_{\hat\Theta}\sup_{\Theta\in\mathcal{B}(\mathcal{Z})}\mathbb{P}\left(\frac{1}{d^3}\|\hat\Theta-\Theta\|_F^2\geq\frac{C\sigma^2\log k}{d^2}\right)=\inf_{\hat\Theta}\sup_{z\in\mathcal{Z}}\mathbb{P}\left(\frac{1}{d^3}\|\hat\Theta-\Theta\|_F^2\geq\frac{C\sigma^2\log k}{d^2}\right)\geq 0.9,$$

with some constant $C>0$ for sufficiently small $c>0$ in (32). $\square$

B.5 Proof of Lemma 2

Proof. Without loss of generality, assume that π is the identity permutation. Notice that $g(i) - \tau(i)$ is the sample average of roughly (excluding repetitions from symmetricity) d^{m-1} independent mean-zero sub-Gaussian random variables with the variance proxy σ. Based on the independence of sub-Gaussian random variables, we have

$$|g(i) - \tau(i)| < 2\sigma d^{-(m-1)/2}\sqrt{\log d}, \tag{39}$$

with probability $1 - \frac{2}{d^2}$ for all $i \in [d]$.

By the weakly β-monotonicity of the function g, we have

$$g(1) \pm \delta \leq g(2) \pm \delta \leq \cdots \leq g(d-1) \pm \delta \leq g(d) \pm \delta, \tag{40}$$

where $\delta \lesssim d^{-(m-1)/2}$ is the small tolerance. The estimated permutation $\hat{\pi}$ is defined for which

$$\tau(\hat{\pi}^{-1}(1)) \leq \tau(\hat{\pi}^{-1}(2)) \leq \cdots \leq \tau(\hat{\pi}^{-1}(d-1)) \leq \tau(\hat{\pi}^{-1}(d)). \tag{41}$$

For any given index i, we examine the error $|i - \hat{\pi}(i)|$. By (40) and (41), we have

$$i = \underbrace{|\{j : g(j) \leq g(i)\}|}_{=:\mathrm{I}}, \quad \text{and} \quad \hat{\pi}(i) = \underbrace{|\{j : \tau(j) \leq \tau(i)\}|}_{=:\mathrm{II}},$$

where $|\cdot|$ denotes the cardinality of the set. We claim that the sets I and II differ only in at most $d^{(m-1)\beta/2}$ elements. To prove this, we partition the indices in $[d]$ in two cases.

1. Long-distance indices in $\{j : |j - i| \geq C\left(\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta}\}$ for some sufficient large constant $C > 0$. In this case, the ordering of (i, j) remains the same in (40) and (41), i.e.,

   $$g(i) < g(j) \Longleftrightarrow \tau(i) < \tau(j). \tag{42}$$

   We only prove the right side direction in (42) here. The other direction can be similarly proved. Suppose that $g(i) < g(j)$. Then we have

   $$\begin{aligned}\tau(j) - \tau(i) &\geq -|g(j) - \tau(j)| - |g(i) - \tau(i)| + g(j) - g(i) \\ &> -4\sigma d^{(m-1)/2}\sqrt{\log d} + g(j) - g(i) \\ &\geq 0,\end{aligned}$$

   where the second inequality is from (39) with probability at least $(1 - 2/d^2)^d$ and the last inequality uses weakly β-monotonousity of $g(\cdot)$, the tolerance condition $\delta \lesssim d^{-(m-1)/2}$, and the assumption $|j - i| \geq C\left(\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta}$. Therefore we show that $g(i) < g(j)$ implies $\tau(i) < \tau(j)$. In this case, we conclude that none of long-distance indices belongs to IΔII.

2. Short-distance indices in $\{j\colon |j-i| < \left(\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta}\}$. In this case, (40) and (41) may yield different ordering of (i,j).

Combining the above two cases gives that

$$\left\{j\colon \frac{1}{d}|j-i| \leq \left(4\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta}\right\} \supset \mathrm{I}\Delta\mathrm{II}.$$

Finally, we have

$$\operatorname{Loss}(\pi,\hat{\pi}) := \frac{1}{d}\max_{i\in[d]}|\pi(i)-\hat{\pi}(i)| \leq \frac{1}{d}\mathrm{I}\Delta\mathrm{II} \leq \left(4\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta},$$

with high probability. □

B.6 Proof of Theorem 3

Proof. By Lemma 1, there exists a block-wise polynomial tensor $\mathcal{B} \in \mathscr{B}(k,\ell)$ satisfying (20). By the triangle inequality, we decompose estimation error into three terms,

$$\begin{aligned}
&\|\hat{\Theta}^{\mathrm{BC}}\circ\hat{\pi}^{\mathrm{BC}} - \Theta\circ\pi\|_F \qquad (43)\\
\leq\ &\|\hat{\Theta}^{\mathrm{BC}}\circ\hat{\pi}^{\mathrm{BC}} - \mathcal{B}\circ\hat{\pi}^{\mathrm{BC}}\|_F + \|\mathcal{B}\circ\hat{\pi}^{\mathrm{BC}} - \Theta\circ\hat{\pi}^{\mathrm{BC}}\|_F + \|\Theta\circ\hat{\pi}^{\mathrm{BC}} - \Theta\circ\pi\|_F\\
=\ &\underbrace{\|\Theta\circ\hat{\pi}^{\mathrm{BC}} - \Theta\circ\pi\|_F}_{\text{Permutation error}} + \underbrace{\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F}_{\text{Nonparametric error}} + \underbrace{\|\mathcal{B}-\Theta\|_F}_{\text{Lemma 1}}.
\end{aligned}$$

Therefore, it suffices to bound two terms $\|\Theta\circ\hat{\pi}^{\mathrm{BC}} - \Theta\circ\pi\|_F$ and $\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F$ separately.

Permutation error. For any $(i_1,\ldots,i_m)\in[d]^m$, we have

$$\begin{aligned}
&\left|\Theta\left(\hat{\pi}^{\mathrm{BC}}(i_1),\ldots,\hat{\pi}^{\mathrm{BC}}(i_m)\right) - \Theta\left(\pi(i_1),\ldots,\pi(i_m)\right)\right|\\
\leq\ &\left\|\left(\frac{\hat{\pi}^{\mathrm{BC}}(i_1)}{d},\ldots,\frac{\hat{\pi}^{\mathrm{BC}}(i_m)}{d}\right) - \left(\frac{\pi(i_1)}{d},\ldots,\frac{\pi(i_m)}{d}\right)\right\|_\infty^{\min(\alpha,1)}\\
\leq\ &\left[\frac{1}{d}\max_{i\in[d]}\left|\hat{\pi}^{\mathrm{BC}}(i)-\pi(i)\right|\right]^{\min(\alpha,1)}\\
\lesssim\ &\left(\sigma d^{-(m-1)/2}\sqrt{\log d}\right)^{\beta\min(\alpha,1)},
\end{aligned}$$

where the first inequality is from the α-Hölder smoothness of Θ, and the last inequality is from Lemma 2. Therefore, we obtain the upper bound of the permutation error

$$\frac{1}{d^m}\|\Theta\circ\hat{\pi}^{\mathrm{BC}} - \Theta\circ\pi\|_F^2 \lesssim \left(\sigma^2\frac{\log d}{d^{m-1}}\right)^{\beta\min(\alpha,1)}. \qquad (44)$$

Nonparametric error. Recall that Borda count estimation is defined by $\hat{\Theta}^{\mathrm{BC}} := \arg\min_{\Theta\in\mathscr{B}(k,\ell)}\|\tilde{\mathcal{Y}}-\Theta\|_F^2$, where $\tilde{\mathcal{Y}} = \mathcal{Y}\circ(\hat{\pi}^{\mathrm{BC}})^{-1}$. By the optimality of least-square estimator, we have

$$
\begin{aligned}
\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F &\leq \left\langle \frac{\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}}{\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F},\ \mathcal{Y} \circ \pi \circ (\hat{\pi}^{\mathrm{BC}})^{-1} - \mathcal{B} \right\rangle \\
&\equiv \left\langle \frac{\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}}{\|\hat{\Theta}^{\mathrm{BC}} - \mathcal{B}\|_F}, \mathcal{E} + \left(\Theta \circ \pi \circ (\hat{\pi}^{\mathrm{BC}})^{-1} - \mathcal{B}\right) \right\rangle \\
&\leq \sup_{\mathcal{B}, \mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B} - \mathcal{B}\|_F}, \mathcal{E} \right\rangle + \|\Theta \circ \pi - \mathcal{B} \circ \hat{\pi}^{\mathrm{BC}}\|_F \\
&\leq \sup_{\mathcal{B}, \mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B} - \mathcal{B}\|_F}, \mathcal{E} \right\rangle + \underbrace{\|\Theta \circ \pi - \Theta \circ \hat{\pi}^{\mathrm{BC}}\|_F}_{\text{Permutation error (44)}} + \underbrace{\|\Theta - \mathcal{B}\|_F}_{\text{Lemma 1}}
\end{aligned}
$$

Now we bound the inner product term. By the same argument in the proof of Theorem 1, the space embedding $\mathscr{B}(k,\ell) \subset \mathbb{R}^{(\ell+m)^\ell k^m}$ implies the space embedding $\{(\mathcal{B}' - \mathcal{B})\colon \mathcal{B}, \mathcal{B}' \in \mathscr{B}(k,\ell)\} \subset \mathbb{R}^{2(\ell+m)^\ell k^m}$. Therefore, we have

$$
\sup_{\mathcal{B}, \mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B}' - \mathcal{B}\|_F}, \mathcal{E} \right\rangle \leq \sup_{\boldsymbol{c} \in \mathbb{R}^{2k^m(\ell+m)^\ell}} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2},\ e \right\rangle, \tag{45}
$$

where $e \in \mathbb{R}^{2k^m(k+m)^\ell}$ is a vector consisting of i.i.d. sub-Gaussian entries with variance proxy σ^2. Combining (45) and Lemma 6 yields

$$
\begin{aligned}
\mathbb{P}\left(\sup_{\mathcal{B}, \mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B}' - \mathcal{B}\|_F}, \mathcal{E} \right\rangle \geq t \right) &\leq \mathbb{P}\left(\sup_{\boldsymbol{c} \in \mathbb{R}^{2k^m(\ell+m)^\ell}} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2},\ e \right\rangle \geq t \right) \\
&\leq \exp\left(-\frac{t^2}{8\sigma^2} + k^m(\ell+m)^\ell \log 6 \right),
\end{aligned}
$$

Setting $t = C\sigma\sqrt{k^m(\ell+m)^\ell}$ for sufficiently large $C > 0$ gives

$$
\sup_{\mathcal{B}, \mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' - \mathcal{B}}{\|\mathcal{B}' - \mathcal{B}\|_F}, \mathcal{E} \right\rangle \lesssim \sigma\sqrt{k^m(\ell+m)^\ell}, \tag{46}
$$

with high probability.

Finally, combining all sources of error from Lemma 1 and inequalities (44), (46), (43) yields

$$
\frac{1}{d^m}\|\hat{\Theta}^{\mathrm{BC}} \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \pi\|_F \lesssim \left(\sigma^2 \frac{\log d}{d^{m-1}}\right)^{\beta \min(\alpha,1)} + \sigma^2 \frac{k^m(\ell+m)^\ell}{d^m} + \frac{L^2}{k^{2\min(\alpha,\ell+1)}}. \tag{47}
$$

Finally, optimizing (47) with respect to (k,l) gives that

$$
(47) \ \lesssim \begin{cases} L^2\left(\frac{\sigma}{L}\right)^{\frac{4\alpha}{m+2\alpha}} d^{-\frac{2m\alpha}{m+2\alpha}}, & \text{when } \alpha < c(\alpha,\beta,m), \\ \left(\frac{\sigma^2 \log d}{d^{m-1}}\right)^{\beta\min(\alpha,1)}, & \text{when } \alpha \geq c(\alpha,\beta,m), \end{cases}
$$

under the choice

$$\ell^* = \min\left(\lfloor \alpha \rfloor, \lfloor c(\alpha,\beta,m) \rfloor\right), \quad k^* = c_1 d^{m/(m+2\min(\alpha,\ell^*+1))},$$

where $c(\alpha,\beta,m) := \frac{m(m-1)\beta\min(\alpha,1)}{\max(0,\ 2(m-(m-1)\beta\min(\alpha,1)))}$.

$\square$

C Technical lemmas

Lemma 5 (Sub-Gaussian maxima under full embedding)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$ be a deterministic matrix with rank $r \le \min(d_1,d_2)$. Let $\boldsymbol{y} \in \mathbb{R}^{d_1}$ be a sub-Gaussian random vector with variance proxy σ^2. Then, there exists a sub-Gaussian random vector $\boldsymbol{x} \in \mathbb{R}^r$ with variance prosy σ^2 such that*

$$\max_{\boldsymbol{p}\in\mathbb{R}^{d_2}} \left\langle \frac{\boldsymbol{A}\boldsymbol{p}}{\|\boldsymbol{A}\boldsymbol{p}\|_2}, \boldsymbol{y} \right\rangle = \max_{\boldsymbol{q}\in\mathbb{R}^{r}} \left\langle \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_2}, \boldsymbol{x} \right\rangle.$$

Proof. Let $\boldsymbol{u}_i \in \mathbb{R}^{d_1}, \boldsymbol{v}_j \in \mathbb{R}^{d_2}$ singular vectors and $\lambda_i \in \mathbb{R}$ be singular values of of $\boldsymbol{A}$ such that $\boldsymbol{A} = \sum_{i=1}^r \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^T$. Then for any $\boldsymbol{p} \in \mathbb{R}^{d_2}$, we have

$$\boldsymbol{A}\boldsymbol{p} = \sum_{i=1}^r \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^T \boldsymbol{p} = \sum_{i=1}^r \lambda_i (\boldsymbol{v}_i^T \boldsymbol{p}) \boldsymbol{u}_i = \sum_{i=1}^r \alpha_i \boldsymbol{u}_i,$$

where $\boldsymbol{\alpha}(\boldsymbol{p}) = (\alpha_1,\dots,\alpha_r)^T := \left(\lambda_1(\boldsymbol{v}_1^T\boldsymbol{p}),\dots,\lambda_r(\boldsymbol{v}_r^T\boldsymbol{p})\right)^T \in \mathbb{R}^r$. Notice that $\boldsymbol{\alpha}(\boldsymbol{p})$ covers $\mathbb{R}^r$ in the sense that $\{\boldsymbol{\alpha}(\boldsymbol{p})\colon \boldsymbol{p} \in \mathbb{R}^{d_2}\} = \mathbb{R}^r$. Therefore, we have

$$\begin{aligned}
\max_{\boldsymbol{p}\in\mathbb{R}^{d_2}} \left\langle \frac{\boldsymbol{A}\boldsymbol{p}}{\|\boldsymbol{A}\boldsymbol{p}\|_2}, \boldsymbol{y} \right\rangle &= \max_{\boldsymbol{p}\in\mathbb{R}^{d_2}} \sum_{i=1}^r \frac{\alpha_i}{\|\boldsymbol{\alpha}(\boldsymbol{p})\|_2} \boldsymbol{u}_i^T \boldsymbol{y} \\
&= \max_{\boldsymbol{p}\in\mathbb{R}^{d_2}} \left\langle \frac{\boldsymbol{\alpha}(\boldsymbol{p})}{\|\boldsymbol{\alpha}(\boldsymbol{p})\|_2}, \boldsymbol{x} \right\rangle \\
&= \max_{\boldsymbol{q}\in\mathbb{R}^{r}} \left\langle \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_2}, \boldsymbol{x} \right\rangle,
\end{aligned}$$

where we define $\boldsymbol{x} = (\boldsymbol{u}_1^T\boldsymbol{y},\dots,\boldsymbol{u}_r^T\boldsymbol{y})^T \in \mathbb{R}^r$. Since $\boldsymbol{u}_i^T\boldsymbol{y}$ is sub-Gaussian with variance proxy σ^2 because of orthonormality of $\boldsymbol{u}_i$, the proof is completed. $\square$

Remark 6. In particular, if $\boldsymbol{x} \in \mathbb{R}^r, \boldsymbol{y} \in \mathbb{R}^{d_1}$ are two Gaussian random vectors with i.i.d. entries drawn from $N(0,\sigma^2)$. Define two Gaussian maximums

$$F(\boldsymbol{x}) \stackrel{\text{def}}{=} \max_{\boldsymbol{q}\in\mathbb{R}^{r}} \left\langle \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_2}, \boldsymbol{x} \right\rangle, \qquad G(\boldsymbol{x}) \stackrel{\text{def}}{=} \max_{\boldsymbol{p}\in\mathbb{R}^{d_2}} \left\langle \frac{\boldsymbol{A}\boldsymbol{p}}{\|\boldsymbol{A}\boldsymbol{p}\|_2}, \boldsymbol{y} \right\rangle.$$

Then $F(\boldsymbol{x}) = G(\boldsymbol{y})$ in distribution. This equality holds because $(\boldsymbol{u}_1^T\boldsymbol{y},\dots,\boldsymbol{u}_r^T\boldsymbol{y})$ is again Gaussian random vectors whose entries are i.i.d. drawn from $N(0,\sigma^2)$.

Lemma 6 (Theorem 1.19 in Phillippe Rigollet [23]). *Let $e \in \mathbb{R}^d$ be a sub-Gaussian vector with variance proxy σ^2. Then,*

$$\mathbb{P}\left(\max_{\boldsymbol{c}\in\mathbb{R}^d}\left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2}, e\right\rangle \geq t\right) \leq \exp\left(-\frac{t^2}{8\sigma^2} + d\log 6\right).$$

Proposition 1 (Proposition 4.1 in Gao et al. [10]). *Let (Ξ, ρ) be a metric space and $\{\mathbb{P}_\xi \colon \xi \in \Xi\}$ be a collection of probability measure. For any totally bounded $T \subset \Xi$, define the Kullback-Leibler diameter of T by $d_{KL}(T) = \sup_{\xi,\xi'\in T} D(\mathbb{P}_\xi|\mathbb{P}_{\xi'})$. Then,*

$$\inf_{\hat{\xi}}\sup_{\xi\in\Xi}\mathbb{P}_\xi\left\{\rho(\hat{\xi},\xi) \geq \frac{\epsilon^2}{4}\right\} \geq 1 - \frac{d_{KL}(T) + \log 2}{\log \mathcal{M}(\epsilon, T, \rho)},$$

where $\mathcal{M}(\epsilon, T, \rho)$ is a packing number of T with respect to the metric ρ.

Lemma 7 (Varshamov-Gilbert bound). *There exists a sequence of subset $\omega_1, \dots, \omega_N \in \{0,1\}^d$ such that*

$$\rho_H(\omega_i, \omega_j) := \|\omega_i - \omega_j\|_F^2 \geq \frac{d}{4} \text{ for any } i \neq j \in [N],$$

for some $N \geq \exp(d/8)$.