Blockwise Polynomial Approximation to Smooth Tensor Model

Miaoyan Wang, Sep 15, 2021

1 Results

For notational convenience, we make the convention that blockwise constant tensor is of degree 1 (not 0 as in classical conventions). We use $z\colon [d] \to [k]$ to denote the canonical clustering function that partitions $[d]$ into k equal-sized clusters such that $z(i) = \lceil ki/d \rceil$. By construction, the inverse images $\{z^{-1}(j)\colon j \in [k]\}$ is a collection of disjoint, equal-sized subsets satisfying $\cup_{j\in[k]} z^{-1}(j) = [d]$. We use $\mathcal{E}_k$ to denote the m-way partition that collects k^m disjoint, equal-sized blocks in $[d]^m$; i.e.,

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

- blockwise degree-1 (constant) tensor:

$$\begin{aligned}\mathscr{B}(k,1) &= \left\{ \mathcal{B} \in (\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(\omega) = \sum_{\Delta\in\mathcal{E}_k} c_\Delta \mathbb{1}\{\omega \in \Delta\} \right\} \\ &\cong \mathbb{R}^{k^m},\end{aligned}$$

  where, for each block $\Delta \in \mathcal{E}_k$, the coefficients $c_\Delta \in \mathbb{R}$ represent the block means. Note that there are in total k^m free parameters in $\mathscr{B}(k,1)$, so the parameter space $\mathscr{B}(k,1)$ is isomorphic to the linear space $\mathbb{R}^{k^m}$.

- blockwise degree-2 linear tensor:

$$\begin{aligned}\mathscr{B}(k,2) &= \left\{ \mathcal{B} \in (\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(\omega) = \sum_{\Delta\in\mathcal{E}_k} [c_\Delta + \langle \boldsymbol{\beta}_\Delta, \omega\rangle]\, \mathbb{1}\{\omega \in \Delta\} \text{ for all indices } \omega \in [d]^m \right\} \\ &\cong \mathbb{R}^{(1+m)k^m},\end{aligned}$$

  where, for each block $\Delta \in \mathcal{E}_k$, the coefficients $(c_\Delta, \boldsymbol{\beta}_\Delta) \in \mathbb{R} \times \mathbb{R}^d$ represent the means and coordinate-wise slopes within blocks. Note that there are in total k^m blocks in $\mathcal{E}_k$, each of which is associated with R^{1+d} free coefficients. By the same argument as before, the parameter space $\mathscr{B}(k,2)$ is isomorphic to the linear space $\mathbb{R}^{(1+m)k^m}$.

- blockwise degree-$(\ell+1)$ polynomial tensor:

$$\begin{aligned}\mathscr{B}(k,\ell+1) &= \left\{ \mathcal{B} \in (\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(\omega) = \sum_{\Delta\in\mathcal{E}_k} \mathrm{Poly}_{\ell,\Delta}(\omega) \mathbb{1}\{\omega \in \Delta\} \text{ for all indices } \omega \in [d]^m \right\} \\ &\subset \mathbb{R}^{(\ell+m)^\ell k^m},\end{aligned}$$

where, for each block $\Delta \in \mathcal{E}_k$, the polynomial function $\mathrm{Poly}_{\ell,\Delta}(\cdot)$ has at most $(\ell+m)^\ell$ free coefficients. By the same argument as before, the parameter space $\mathscr{B}(k, \ell+1)$ is embedded in the linear space $\mathbb{R}^{(\ell+m)^\ell k^m}$.

Model. Suppose the data tensor $\mathcal{Y}$ is generated from the model

$$\mathcal{Y} = \Theta \circ \pi + \mathcal{E}, \quad \text{where} \quad \Theta(i_1, \ldots, i_m) = f\left(\frac{i_1}{d}, \ldots, \frac{i_m}{d}\right) \text{ for all } (i_1, \ldots, i_d) \in [d]^m, \tag{1}$$

where $\pi\colon [d] \to [d]$ is an *unknown* permutation, $f\colon \mathbb{R}^m \to \mathbb{R}$ is an *unknown* α-Hölder smooth function with $\alpha \in (0, \infty)$, and $\mathcal{E}$ is a noise tensor with i.i.d. sub-Gaussian entries. We use $\mathcal{P}(\alpha)$ to denote the collection of signal tensors from model (1). The goal is to estimate signal $\Theta \in \mathcal{P}(\alpha)$ from data $\mathcal{Y}$.

The parameters (Θ, π) are not separately identifiable from model (1). However, the tensor $\Theta \circ \pi$ is always identifiable as a composite parameter. We impose the following marginal monotonicity assumption to ensure the separate identifiability.

Theorem 1 (Identifiability)**.** Suppose $f \in \mathcal{M}(\beta)$ with $\beta \in (0, \infty)$. Then, the parameters (Θ, π) are separately identifiable from model (1).

Theorem 2 (Blockwise polynomial tensor approximation)**.** Suppose the function $f\colon [0,1]^m \to \mathbb{R}$ generating the signal tensor Θ is α-Hölder smooth with $\alpha \in (0, \infty)$. Then, for every block size $k \le d$ and degree $\ell \in \mathbb{N}_+$, we have the approximation error

$$\inf_{\mathcal{B} \in \mathscr{B}(k,\ell)} \frac{1}{d^m} \|\Theta - \mathcal{B}\|_F^2 \lesssim \frac{m^2}{k^{2\min(\alpha,\ell)}}.$$

We propose a least-square estimate based on the blockwise polynomial tensor approximation,

$$(\hat{\Theta}^{\mathrm{LSE}}, \hat{\pi}^{\mathrm{LSE}}) = \mathop{\arg\min}_{\substack{\Theta \in \mathscr{B}(k,\ell) \\ \pi\colon [d] \to [d]}} \|\mathcal{Y} - \Theta \circ \pi\|_F^2.$$

Although not reflected in the notation, the least-square estimate $\hat{\Theta}^{\mathrm{LSE}}$ depends on the tuning parameters (k, ℓ). We provide the optimal choice of (k, ℓ) in the following theorem. We focus on the asymptotic error rates as $d \to \infty$ while treating (m, α) as constants.

Theorem 3 (Least-square estimator)**.** Let $(\hat{\Theta}^{\mathrm{LSE}}, \hat{\pi}^{\mathrm{LSE}})$ denote the least-square estimate with degree $\ell^* = \min(\lceil \alpha \rceil, \frac{m(m-1)}{2})$ with block size $k^* = \lceil d^{\frac{m}{m+2\ell^*}} \rceil$. Then, $(\hat{\Theta}^{\mathrm{LSE}}, \hat{\pi}^{\mathrm{LSE}})$ obeys

$$\begin{aligned}
\frac{1}{d^m} \|\hat{\Theta}^{\mathrm{LSE}} \circ \hat{\pi}^{\mathrm{LSE}} - \Theta \circ \pi\|_F^2 &\lesssim \inf_{(k,\ell) \in [d] \times \mathbb{N}_+} \left\{ \frac{m^2}{k^{2\min(\alpha,\ell)}} + \frac{k^m (\ell+m)^\ell}{d^m} + \frac{\log d}{d^{m-1}} \right\} \\
&\asymp \begin{cases} d^{-\frac{2m\alpha}{m+2\alpha}} & \text{when } \alpha < m(m-1)/2, \\ d^{-(m-1)} \log d & \text{when } \alpha \ge m(m-1)/2. \end{cases}
\end{aligned}$$

Remark 1 (Comparison with block tensor approximation)**.** For matrices (i.e., $m = 2$), the optimal polynomial is obtained by block matrix approximation. For order-3 α-smooth tensors the optimal degree and block size are $(\ell^*, k^*) = (3, \lceil d^{1/3} \rceil)$ for all $\alpha \geq 3$. In other words, blockwise quadratic tensors suffice for estimating sufficiently smooth tensors. Further increment of polynomial degree ℓ is of no help for smooth signal estimation.

Theorem 4 (Polynomial-time estimator)**.** Suppose that the signal tensor Θ is generated from model (1) with $f \in \mathcal{H}(\alpha) \cap \mathcal{M}(\beta)$. Let $\hat{\Theta}^{\mathrm{BC}}$ be the estimator in with degree $\ell^* = \min(\lceil \alpha \rceil, \frac{m(m-1)}{2})$ and block size $k^* = \lceil d \frac{m}{m+2\ell^*} \rceil$. Then the estimator $\hat{\Theta}^{\mathrm{BC}}$ satisfies

$$\frac{1}{d^m} \|\hat{\Theta}^{\mathrm{BC}} \circ \hat{\pi}^{\mathrm{BC}} - \Theta \circ \pi\|_F^2 \lesssim d^{-\beta(m-1)} + \begin{cases} d^{-\frac{2m\alpha}{m+2\alpha}} & \text{when } \alpha < m(m-1)/2, \\ d^{-(m-1)} \log d & \text{when } \alpha \geq m(m-1)/2. \end{cases}$$

with very high probability.

Theorem 5 (Minimax lower bound)**.** For any given $\alpha \in (0, \infty)$, the estimation problem based on model (1) obeys the minimax lower bound

$$\inf_{(\hat{\Theta}, \hat{\pi})} \sup_{\substack{\Theta \in \mathcal{P}(\alpha) \\ \pi \colon [d] \to [d]}} \mathbb{P}\left(\frac{1}{d^m} \|\Theta \circ \pi - \hat{\Theta} \circ \hat{\pi}\|_F^2 \geq d^{-\frac{2m\alpha}{m+2\alpha}} + d^{-(m-1)} \log d \right) \geq 0.8. \tag{2}$$

Remark 2. By comparing Theorems 3 and 5, we find that the constrained least-square estimator achieves the minimax optimal rate.

2 Proofs of Main Theorems

Proof of Theorem 3. The proof is similar to theorem 2.1 on note 030721. By Theorem 2, there exists a blockwise polynomial tensor $\mathcal{B} \in \mathscr{B}(k, \ell)$ such that

$$\|\mathcal{B} - \Theta\|_F^2 \lesssim \frac{d^m m^2}{k^{2\min(\alpha, \ell)}}. \tag{3}$$

By the triangle inequality,

$$\|\hat{\Theta}^{\mathrm{LSE}} \circ \hat{\pi}^{\mathrm{LSE}} - \Theta \circ \pi\|_F^2 \leq 2\|\hat{\Theta}^{\mathrm{LSE}} \circ \hat{\pi}^{\mathrm{LSE}} - \mathcal{B} \circ \pi\|_F^2 + 2 \underbrace{\|\mathcal{B} \circ \pi - \Theta \circ \pi\|_F^2}_{\text{Theorem 2}}. \tag{4}$$

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

$$\|\hat{\Theta}^{\mathrm{LSE}} \circ \hat{\pi}^{\mathrm{LSE}} - \mathcal{B} \circ \pi\|_F \leq \left\langle \frac{\hat{\Theta}^{\mathrm{LSE}} \circ \hat{\pi}^{\mathrm{LSE}} - \mathcal{B} \circ \pi}{\|\hat{\Theta}^{\mathrm{LSE}} \circ \hat{\pi}^{\mathrm{LSE}} - \mathcal{B} \circ \pi\|_F}, \ \mathcal{E} + (\mathcal{B} \circ \pi - \Theta \circ \pi) \right\rangle$$

$$\leq \sup_{\pi,\pi'\colon [d]\to[d]} \sup_{\mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}'\circ\pi' - \mathcal{B}\circ\pi}{\|\mathcal{B}'\circ\pi' - \mathcal{B}\circ\pi\|_F}, \mathcal{E} \right\rangle + \underbrace{\|\mathcal{B}\circ\pi - \Theta\circ\pi\|_F}_{\text{Theorem 2}}.$$

Now, for fixed π, π', the space embedding $\mathscr{B}(k,\ell) \subset \mathbb{R}^{(\ell+m)^\ell k^m}$ implies the space embedding $\{(\mathcal{B}'\circ\pi' - \mathcal{B}\circ\pi)\colon \mathcal{B},\mathcal{B}'\in\mathscr{B}(k,\ell)\} \subset \mathbb{R}^{2(\ell+m)^\ell k^m}$. Therefore, with very high probability,

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

where e is a vector of consistent length that consists of i.i.d. sub-Gaussian entries. By the union bound of Gaussian maxima over countable set $\{\pi, \pi'\colon [d]\to[d]\}$, we obtain

$$\mathbb{E}\|\hat{\Theta}^{\mathrm{LSE}}\circ\hat{\pi}^{\mathrm{LSE}} - \mathcal{B}\circ\pi\|_F^2 \lesssim (\ell+m)^\ell k^m + d\log d. \tag{5}$$

Combining the inequalities (3), (4) and (5) yields the desired conclusion

$$\mathbb{E}\|\hat{\Theta}^{\mathrm{LSE}}\circ\hat{\pi}^{\mathrm{LSE}} - \Theta\circ\pi\|_F^2 \lesssim \frac{d^m m^2}{k^{2\min(\alpha,\ell)}} + (\ell+m)^\ell k^m + d\log d.$$

□

Proof of Theorem 5. By the definition of the tensor space, we seek the minimax rate ε^2 in the following expression

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

On one hand, if we fix permutation $\pi\colon [d]\to[d]$, 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 $\varepsilon^2 = d^{-\frac{2m\alpha}{m+2\alpha}}$. On the other hand, if we fix $\Theta\in\mathcal{P}(\alpha)$, the problem become a new type of convergence rate due to the unknown permutation. We refer it to the permutation rate, and will prove that $\varepsilon^2 = d^{-(m-1)}\log d$. Since our target is the sum of the two rate, 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 Wasserman [2019, Lecture note, example 16] and Stone [1982, Section 2]. We state the results here for self-completeness.

Lemma 6 (Minimax rate for α-smooth function estimation)**.** Consider data $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_N, Y_N)$, where $\boldsymbol{x}_n = (\frac{i_1}{d}, \ldots, \frac{i_m}{d}) \in [0,1]^d$ is the m-dimensional predictor and $Y_n\in\mathbb{R}$ is the scalar response. Consider the observation model

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

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

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}(\alpha)} \mathbb{P}\left(\|f - \hat{f}\|_2 \geq N^{-\frac{2\alpha}{m+2\alpha}} \right) \geq 0.9. \tag{6}$$

Our desired nonparametric rate readily follows from Lemma 6 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 $\pi \in [d] \to [d]$, we have

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

Permutation rate. Let $\Pi(n, k)$ denote the collection of all possible onto mappings from $[n]$ to $[k]$. Because our main focus in the unknown permutation, we define the following two (conditional) tensor families by fixing the generative function and the core tensor, respectively.

- d-dimensional, α-smooth tensor family with a given α-smooth function f:

$$\mathcal{Q}(d, \alpha | f) = \left\{ \Theta \colon \Theta(i_1, \ldots, i_m) = f\left(\frac{\pi(i_1)}{d}, \ldots, \frac{\pi(i_m)}{d} \right) \text{ for some onto mapping } \pi \in \Pi(d, d) \right\}.$$

- n-dimensional, k-block tensor family with a given block mean tensor $\mathcal{C}$:

$$\mathcal{Z}(n, k | \mathcal{C}) = \left\{ \Theta \colon \Theta(i_1, \ldots, i_m) = \mathcal{C}(\pi(i_1), \ldots, \pi(i_m)) \text{ and some onto mapping } \pi \in \Pi(n, k) \right\}.$$

The permutation rate is obtained by the two steps. We first show that the permutation rate for the tensor block problem is $\tilde{\mathcal{O}}(d)$. Then, we show that the problem of permutation estimation for smooth tensors is no easier than that for block tensors. Taken together, we establish the permutation rate for smooth tensor models. We provide two key lemmas to show the above two steps respectively; their proofs are deferred to Appendix.

Lemma 7 shows the permutation rate over parameter space $\mathcal{Z}(n, k \mid \mathcal{C})$ is $n^{-(m-1)} \log k$.

Lemma 7 (Permutation error for tensor block model)**.** Consider the problem of estimating n-dimensional, k-block signal tensors from sub-Gaussian tensor block models. For every given integer $k \in [n]$, there exists a core tensor $\mathcal{C} \in \mathbb{R}^{k \times \cdots \times k}$ satisfying

$$\inf_{\hat{\Theta}} \sup_{\pi \in \Pi(n,k)} \mathbb{P}\left\{ \frac{1}{n^m} \sum_{(i_1, \ldots, i_m) \in [n]^m} \left[\hat{\Theta}(i_1, \ldots, i_m) - \mathcal{C}\left(\pi(i_1), \ldots, \pi(i_m)\right) \right]^2 \gtrsim n^{-(m-1)} \log k \right\} \geq 0.9. \tag{7}$$

The proof of Lemma 7 is constructive. We will write $\{\mathcal{C}_k \in \mathbb{R}^{k\times\cdots\times k}\}_{k\geq 1}$ the collection of block mean tensors satisfying (7), and use them to construct the smooth tensors.

Definition 1 (Reduciblility)**.** A tensor $\Theta \in \mathbb{R}^{d\times\cdots\times d}$ is called *reducible to* a tensor $\Theta_{\text{sub}} \in \mathbb{R}^{n\times\cdots\times n}$, if and only if, there exists a subset of indices $\mathcal{I} \subset [d]$ with $n = |\mathcal{I}|$, such that

$$\Theta_{\text{sub}} = \Theta(\mathcal{I}, \ldots, \mathcal{I}).$$

Lemma 8 constructs a smooth tensor $\Theta \in \mathcal{Q}(d, \alpha \mid f)$ such that its subtensor is $\Theta_{\text{sub}} \in \mathcal{Z}(n, k \mid \mathcal{C})$.

Lemma 8 (Reducing smooth tensors to block tensors with given block means)**.** Let $\{\mathcal{C}_k\}_{k\geq 1}$ denote a collection of tensors, where $\mathcal{C}_k \in \mathbb{R}^{k\times\cdots\times k}$ is an arbitrarily given tensor of dimension k. For every dimension $d \in \mathbb{N}_+$ and smoothness index $\alpha > 0$, there exist an α-smooth function f and a signal tensor $\Theta \in \mathcal{Q}(d, \alpha \mid f)$, such that Θ is reducible to $\Theta_{\text{sub}} \in \mathcal{Z}(n, k \mid \mathcal{C}_k)$. Here (n, k) are two integers satisfying $(n, k) = (\gamma_1 d, d^{\gamma_2})$ for two constants $(\gamma_1, \gamma_2) \in (0, 1)^2$.

Note that the Lemma 8 is different from the block tensor approximation in Theorem 2. While both results relate a smooth tensor to a block tensor, the construction in Lemma 8 requires the block mean to be an *arbitrarily given* tensor $\mathcal{C}$. We address this constraint by constructing a smooth tensor of a slightly larger dimension $d = n/\gamma_1$ with $\gamma_1 > 0$. The asymptotical equivalence $d \asymp n$ suffices for our purpose.

Now, we are ready to prove the permutation rate. Let $\{\mathcal{C}_k\}_{k\geq 1}$ denote the tensors satisfying (7) in Lemma 7. By Lemma 8, we can construct an α-smooth tensor Θ and its subtensor Θ_{sub} that satisfy

$$\Theta_{\text{sub}} \in \mathcal{Z}(n, k \mid \mathcal{C}_k) \text{ for some } n \asymp d \text{ and } k \asymp d^{\gamma_2} \text{ with } \gamma_2 > 0.$$

Based on this particular smooth tensor Θ, we have

$$\begin{aligned}
&\inf_{(\hat{\Theta},\hat{\pi})} \sup_{\pi\in\Pi(d,d)} \mathbb{P}\left(\frac{1}{d^m}\|\hat{\Theta}\circ\hat{\pi} - \Theta\circ\pi\|_F^2 \geq \varepsilon^2\right) \\
&= \inf_{\hat{\Theta}} \sup_{\pi\in\Pi(d,d)} \mathbb{P}\left(\frac{1}{d^m}\|\hat{\Theta} - \Theta\circ\pi\|_F^2 \geq \varepsilon^2\right) \\
&\geq \inf_{\hat{\Theta}_{\text{sub}}} \sup_{\pi_{\text{sub}}\in\Pi(n,n)} \mathbb{P}\left(\frac{1}{n^m}\|\hat{\Theta}_{\text{sub}} - \Theta_{\text{sub}}\circ\pi_{\text{sub}}\|_F^2 \geq \frac{d^m}{n^m}\varepsilon^2\right) \\
&\gtrsim \inf_{\hat{\Theta}_{\text{sub}}} \sup_{\bar{\pi}_{\text{sub}}\in\Pi(n,k)} \mathbb{P}\left(\frac{1}{n^m}\sum_{(i_1,\ldots,i_m)\in[n]^m}\left[\hat{\Theta}_{\text{sub}}(i_1,\ldots,i_m) - \mathcal{C}(\bar{\pi}_{\text{sub}}(i_1),\ldots,\bar{\pi}_{\text{sub}}(i_m)\right]^2 \gtrsim \varepsilon^2\right), \quad (8)
\end{aligned}$$

where the first inequality follows from the reducibility of Θ to Θ_{sub}, and the second inequality follows from the properties $\Theta_{\text{sub}} \in \mathcal{Z}(n, k \mid \mathcal{C})$ and $n \asymp d$. We conclude the desired probability

lower bound by taking $\varepsilon^2 = d \log k$ in combination with Lemma 7 and the property $k \asymp d^{\gamma_2}$. In summary, there exists a smooth tensor $\Theta \in \mathcal{P}(\alpha)$ such that

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

Combining two rates. Finally, we combine (6) and (8) to get the desired lower bound. For any $\Theta, \hat{\Theta} \in \mathcal{P}(\alpha)$, by union bound, we have

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

Taking sup on both sides with the fact

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

yields the desired rate (2). □

3 Proof Skecthes of Lemmas

We provide the proof for $m = 3$ only. The extension to higher orders are similar and omitted.

Proof sketch of Lemma 7. The proof extends Gao et al. [2015, Theorem 2.2, page 26] from matrices to tensors. For the matrix $\boldsymbol{B} \in \mathbb{R}^{k/2 \times k/2}$ defined in Gao et al. [2015, Theorem 2.2, page], we define an order-3 symmetric tensor $\mathcal{C} \in \mathbb{R}^{k \times k \times k}$ by

$$\mathcal{C}(:,:,1) = \mathcal{C}(:,1,:) = \mathcal{C}(1,:,:) = \begin{bmatrix} 0 & \boldsymbol{B} \\ \boldsymbol{B}^T & 0 \end{bmatrix}.$$

Following the same calculation as in Gao et al. [2015], we can verify the tensor $\mathcal{C}$ satisfies (7). □

Proof sketch of Lemma 8. The construction follows the same line as in Gao et al. [2015, Supplement, page 3]. For the same ϕ function defined in Gao et al. [2015], we define a 3-variable function

$$f(x,y,z) = \sum_{a,b,c \in [k]} \left(\mathcal{C}(a,b,c) - \frac{1}{2} \right) \phi\left(kx - a + \frac{1}{2} \right) \phi\left(kx - b + \frac{1}{2} \right) + \phi\left(kx - c + \frac{1}{2} \right) + \frac{1}{2}.$$

Then, it is easy to verify that (1) f is an α-smooth tensor when $k = d^{\gamma_2}$, (2) f has piecewise-constant

shape when restricted to a sub-domain,

$$f(x, y, z) = \mathcal{C}(a, b, c), \quad \text{when} \quad (x, y, z) \in I \times I \times I,$$

for some $I \subset [0, 1]$ and $|I| \geq \gamma_1 > 0$. Therefore, the proof is complete. □

References

Larry Wasserman. Minimax theory. *Lecture notes, http://www.stat.cmu.edu/~larry/=sml/*, 2019.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

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