High-Rank Tensor Estimation Under Smooth Single Index Models

Chanwoo Lee CHANWOO.LEE@WISC.EDU
Miaoyan Wang MIAOYAN.WANG@WISC.EDU
Department of Statistics, University of Wisconsin - Madison

Abstract

We study the semi-parametric single index model with a near low-rank regression coefficient matrix, which extends notion of linear trace regression. Specifically, given matrix covariate $\mathcal{Y} = f(\boldsymbol{B}) + \mathcal{E}$ with $f\colon \mathbb{R} \to \mathbb{R}$ an *unkonwn* analytic function, and $\boldsymbol{B}$ is an unknown low-rank coefficient matrix. This model accommodates various types of responses and embraces many important problem setups such as reduced-rank regression, matrix-response regression, 1-bit matrix completion and compressed sensing among others. We quantify the spectral decay of $\mathbb{E}(\mathcal{Y})$, develop a higher-order spectral algorithm, and show the distinct behavior between the matrix and tensor version of the problem. We first establish a general theory and then for each specific problem, we derive explicitly the statistical rate of the proposed estimator. They all match the minimax rates in the nonparametric regression up to logarithmic factors. Numerical studies confirm the rates we established and demonstrate the advantage of generalized trace regression over linear trace regression when the response is dichotomous. We also show the benefit of incorporating nuclear norm regularization in dynamic stock return prediction and in image classification.

Keywords: Nonparametric regression, matrix completion, Holder smoothness

1. Introduction

We use $f^{(\ell)}(x)$ to denote the ℓ-th derivative of f, evaluated at x. We use $\lesssim$ to denote two sequence up to a universal constant C.

2. Model

Definition 1 (Analytic function class) *Let $C > 0$ be a positive constant. The analytic function class $\mathscr{F}(C)$ on $[-1,1]$ is defined as the set of functions $f\colon [-1,1] \to \mathbb{R}$ whose derivatives satisfy*

$$\sup_{x\in[-1,1]} |f^{(\ell)}(x)| \le C^{\ell+1}\ell! \quad \textit{for all } \ell \in \mathbb{N}_+.$$

Equivalently, f is infinitely differentiable and its Taylor expansion around any point in its domain converges to the function.

A higher-order tensor $\mathcal{T}$ can be unfolded into a matrix. We now introduce several quantities that controls the complexity of matrix unfolding. We use $\text{Mat}(\mathcal{T})$ to denote the matrix unfolding. We use $\text{Trank}(\cdot)$ and $\lambda(\mathcal{T})$ the rank and spectral norm of the matrix $\text{Mat}(\mathcal{T})$,

$$\text{Trank}(\mathcal{T}) := \text{rank}(\text{Mat}(\mathcal{T})), \quad \lambda(\mathcal{T}) := \|\text{Mat}(\mathcal{T})\|_{\text{sp}}.$$

The following family consists of d-dimensional order-m tensor with Tucker rank bounded by r.

Definition 2 (Low-rank tensor class) *The family of d-dimensional rank-s tensor $\mathscr{T}(d, s, m)$ is defined as the set of tensors with Tucker rank bounded by r:*

$$\mathscr{T}(d, s, m) = \{\mathcal{T} \in (\mathbb{R}^d)^{\otimes m} \colon \text{Trank}(\mathcal{T}) \leq s \text{ and } \lambda(\mathcal{T}) \leq 1\}$$

Equivalently, the tensor in class $\mathscr{T}(d, s, m)$ admits the rank-s Tucker decomposition:

$$\mathcal{T} = \mathcal{C} \times_1 \boldsymbol{X} \times \cdots \times_k \boldsymbol{X}.$$

where $\mathcal{C} \in \mathbb{R}^{s \times \cdots \times s}$ is core tensor and $\boldsymbol{X}$ are factor matrices with orthornormal columns. The condition $\lambda(\mathcal{T}) \leq 1$ is imposed without loss of generality. The scale between $\mathcal{T}$ and the f is undetermined; the tensor $f(\mathcal{T}) = f'(\mathcal{T}')$ are the same by setting $\mathcal{T}' = c\mathcal{T}$ and $f' = f/c$. We also note that, the rank-s CP tensor is automatically included in $\mathscr{T}(d, s, m)$.

Now, we are ready to describe the main model. Let $\mathcal{Y}$ be the data tensor. We propose the following observation model

$$\mathcal{Y} = f(\mathcal{T}) + \mathcal{E}, \quad \text{for some unknown } \mathcal{T} \in \mathscr{T}(d, s, m) \text{ and } f \in \mathscr{F}(C), \tag{1}$$

where we assume the noise tensor $\mathcal{E}$ consists of i.i.d. entries with zero-mean and sub-Gaussian parameter σ^2. We call (1) the *single index tensor model*. The name of single index model comes from the observation that

$$\mathbb{E}\left[\mathcal{Y}(\omega) | \mathcal{X}(\omega)\right] = f(\langle \mathcal{T}, \mathcal{X}(\omega) \rangle), \quad \text{for all } \omega \in [d]^m,$$

where, for every index ω, the predictor $\mathcal{X}(\omega) \in (\mathbb{R}^d)^{\otimes m}$ is a dummy-variable tensor with 1 at the ω-th position, and zero everywhere else. Note that the signal tensor $f(\mathcal{T})$ is often high rank. Our goal is to address the following two questions:

- What are the *statistical* and *computational* limits for signal estimation in single index model?
- Are there any intrinsic distinctions for matrices $m = 2$ vs. tensors $m \geq 3$ for high-rank estimation based on model (1)?

3. Identifiability

f and $\mathcal{T}$ are not identifiable separately. However, $f(\mathcal{T})$ is identifiable (why??).

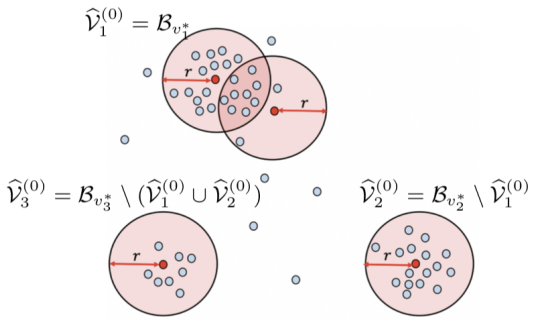
4. Smooth single index models are of log-rank

Let $\mathcal{T} \in \mathscr{T}(d, s, m)$, and $\mathcal{T}^{\circ \ell}$ be the tensor of the same size, with entrywise polynomial transformation $a \to a^\ell$. We have the following rank bound.

Proposition 1 (Monomial tensor) *For every tensor $\mathcal{T} \in \mathscr{T}(d, s, m)$ and every natural number $\ell \in \mathbb{N}_+$*

$$\text{Trank}(\mathcal{T}^{\circ \ell}) \leq \binom{\ell + s - 1}{s - 1}.$$

This proposition shows the polynomial rank growth with respect to ℓ. The exponent depends on the original rank s, which is assumed small. When $s = 1$, then $\text{Trank}(\mathcal{T}^{\circ \ell}) = 1$ for all $\ell \in \mathbb{N}_+$. The bound is nontrivial when $\ell \leq d$. In fact, later we will set $\ell \lesssim \log d$.



Proposition 2 (Uniform approximation for single index tensor) *Consider $\mathcal{T} \in \mathscr{T}(d, s, m)$ and $f \in \mathscr{F}(C)$. There exists a set of basis tensors $\{\mathcal{B}_{\ell,k} \colon (k, \ell) \in [d] \times \mathbb{N}\}$, such that, for every number of pieces $k \in [d]$ and degree $\ell \in \mathbb{N}$, we have*

$$\|f(\mathcal{T}) - \mathcal{B}_{k,\ell}\|_\infty \le C \left(\frac{k}{sm}\right)^{-(\ell+1)} \quad \textit{and} \quad \mathrm{Trank}(\mathcal{B}_{k,\ell}) \le k^s \ell^s.$$

The following is the key property of single index matrix by taking $k \gtrsim \Omega(sm)$ and $\ell = r^{1/s} k^{-1}$ in Proposition 2.

Theorem 4.1 (Dimension-free approximation error) *Consider a single index tensor $\Theta = f(\mathcal{T})$, where $\mathcal{T} \in \mathscr{T}(d, s, m)$ and $f \in \mathscr{F}(C)$. For every rank $r \in \mathbb{N}_+$, we have*

$$\frac{1}{d^m} \|\Theta - \mathrm{Proj}_r(\Theta)\|_F^2 \lesssim C^2 \exp\left(-\frac{r^{1/s}}{sm}\right).$$

The bound is uniform over $\mathcal{T} \in \mathscr{T}(d, s, m)$ and $f \in \mathscr{F}(C)$.

Corollary 1 (Nice smooth tensor model is of $\log$ rank) *For any fixed $\varepsilon > 0$, we have*

$$\mathrm{Trank}_\varepsilon(\Theta) := \min\{\mathrm{Trank}(\mathcal{A}) \colon \|\mathcal{A} - \Theta\|_\infty \le \varepsilon\} \lesssim \log^s d.$$

Corollary 2 (Block tensor model with diverging number of blocks) *Let Θ be a block tensor with s blocks on each mode. The ε-rank of Θ is no more than $r = \log d$.*

When $f(x) = x$, $\mathrm{Trank}(\Theta) = o(1)$. When oscillating $f(x) = \sin(x)$, ...? When monotonic $f(x) = \frac{\exp(-x)}{1+\exp(-x)}$ at bounded regime $[a, b]$; f is quantile transformation??

Why do we define domain on $[-1, 1]$??

Marginal quantile transformed tensor estimation. Low rank. Order-preserving Transformed. Analogy of distance matrix???

Connection to sign-representable tensor. Require f to be monotonic. Here smooth f suffices...

5. Single index model estimation for matrices

5.1. Spectral method and matrix lasso are both nearly minimax

We now consider two estimations. The first one is convex estimater

$$\hat{\mathbf{\Theta}} = \arg\min_{\Theta} \left\{ \|\boldsymbol{Y} - \mathbf{\Theta}\|_F^2 + \sqrt{d}\sigma\|\mathbf{\Theta}\|_* \right\},$$

and the other is non-convex estimator $r = sm\log^s d$. The closed-form solution is

$$\hat{\mathbf{\Theta}} = \sum_{i=1}^{d} (\lambda_i(\boldsymbol{Y}) - \sqrt{d}\sigma)_+ \mathbf{u}_i^T \mathbf{v}_j.$$

5.2. Nonconvex

Theorem 5.1 (Polynomial estimation of single index matrix) *Let $\mathbf{\Theta} = f(\boldsymbol{T})$ with $\boldsymbol{T} \in \mathscr{T}(d, s, 2)$ and $f \in \mathscr{F}(C)$.*

$$\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F \lesssim d^{-1}\log^s(d).$$

Theorem 5.2 (Minimax for matrices) *Assume $s \asymp \mathcal{O}(1)$. Denote the class $\mathscr{F} \circ \mathscr{T} = \{\Theta : \Theta = f(\boldsymbol{T})$ for some $(\boldsymbol{T}, f) \in \mathscr{T}(d, s, 2) \times \mathscr{F}(C)\}$. Then, when we have*

$$\inf_{\hat{\Theta}} \sup_{\mathbf{\Theta} \in \mathscr{F}\circ\mathscr{T}} \mathbb{P}\left(\mathcal{R}(\mathbf{\Theta}, \hat{\mathbf{\Theta}}) \gtrsim d^{-1} \right) \geq 0.2.$$

Proof For every $\alpha_1 \leq \alpha_2 \in (0, \infty)$

$$\mathcal{P}(d, \alpha_1, s, L) \supset \mathcal{P}(d, \alpha_2, s, L) \supset \cdots \mathcal{P}(d, \infty, s, L) \supset \{\Theta\colon \text{Trank}(\Theta) \leq s\}.$$

■

Open question: Can we show sharper lower bound for $\alpha \neq \infty$? Graphon???

6. Tensor estimation

6.1. Non-convex double spectral algorithm

Theorem 6.1 (Polynomial algorithm) *low-rank approximation is substantially better.*

$$\mathcal{R}(\hat{\mathbf{\Theta}}, \mathbf{\Theta}) \leq \begin{cases} r^m + d^{m/2}r + d^m r^{-2\alpha} & r^{-2\alpha} \leq d^{-(m-1)}, \\ \min(d^{-\frac{m}{2}+\frac{(m-\alpha-1)m}{2\alpha}}, 1) & \alpha \leq \frac{(m-1)^2}{m},\ r \asymp \min(d^{(m-1)/2\alpha}, d) \\ d^{-\frac{m}{2}+\frac{m-1}{2\alpha}} & \alpha \geq \frac{(m-1)^2}{m},\ r \asymp d^{\frac{m-1}{2\alpha}} \leq d^{\frac{m}{2(m-1)}} \\ d^{-\frac{m}{2}}\log d & \alpha = \infty,\ r \asymp \log d \end{cases} \tag{2}$$

Block approximation (bottleneck: truncate at $\alpha = 1$):

$$\mathcal{R}(\hat{\mathbf{\Theta}}, \mathbf{\Theta}) \leq \begin{cases} d^{-\frac{2\alpha m}{m+2\alpha}} & \alpha \leq 1 \\ d^{-\frac{2m}{m+2}} & \alpha \geq 1 \end{cases} \tag{3}$$

Theorem 6.2 (Minimax optimality for tensors) *Order-m with known design (bottleneck: truncate at $d^{-(m-1)}$ with unknown design)*

$$\mathcal{R}(\hat{\mathbf{\Theta}}, \mathbf{\Theta}) \leq \begin{cases} d^{-\frac{2\alpha m}{m+2\alpha}} & \alpha \leq \frac{m(m-1)}{2} \\ d^{-(m-1)} \log d & \alpha \geq \frac{m(m-1)}{2} \end{cases} \tag{4}$$

6.2. Statistical and Computational Trade-offs

Acknowledgments

Funding here.

Can we apply tensor algorithm to matrices??

7. Nice tensor model has $\log d$ rank

Appendix A. Proof of Proposition 1

Proof By definition $\mathcal{T} \in \mathscr{T}(d, s, m)$, $\text{Trank}(\text{Mat}(\mathcal{T})) \leq r$. Therefore, there exists matrix SVD such that

$$\text{Mat}(\mathcal{T}) = \sum_{i \in [s]} \lambda_i \boldsymbol{a}_i \otimes \boldsymbol{b}_i,$$

where $\boldsymbol{a}_i \in \mathbb{R}^d$, $\boldsymbol{b} \in \mathbb{R}^{d^{m-1}}$, and $\lambda_1 \geq \cdots \lambda_s \geq 0$. By definition

$$\text{Mat}(\mathcal{T}^{\circ \ell}) = [\text{Mat}(\mathcal{T})]^{\circ \ell} = \left(\sum_{i \in [s]} \lambda_i \boldsymbol{a}_i \otimes \boldsymbol{b}_i \right)^{\circ \ell} \tag{5}$$

$$= \sum_{\substack{\kappa_1 + \cdots + \kappa_s = \ell, \\ (\kappa_1, \ldots, \kappa_s) \in \mathbb{N}_+^s}} \lambda_1^{\kappa_1} \cdots \lambda_s^{\kappa_s} (\boldsymbol{a}_1^{\circ \kappa_1} \circ \cdots \circ \boldsymbol{a}_s^{\circ \kappa_s}) \otimes (\boldsymbol{b}_1^{\circ \kappa_1} \circ \cdots \circ \boldsymbol{b}_s^{\circ \kappa_s}). \tag{6}$$

Here $(\boldsymbol{a}_1^{\circ \kappa_1} \circ \cdots \circ \boldsymbol{a}_s^{\circ \kappa_s}) \in \mathbb{R}^d$ and $(\boldsymbol{b}_1^{\circ \kappa_1} \circ \cdots \circ \boldsymbol{b}_s^{\circ \kappa_s}) \in \mathbb{R}^{d-1}$. Now notice that by counting argument,

$$\#\{(\kappa_1, \ldots, \kappa_s) \in \mathbb{N}_+^s : \kappa_1 + \cdots + \kappa_s = \ell\} = \binom{\ell + s - 1}{s - 1}.$$

Therefore, the summation (5) consists of no more than $\binom{\ell+s-1}{s-1}$ rank-1 terms. We conclude that

$$\text{Trank}(\mathcal{T}^{\circ \ell}) = \text{rank}(\text{Mat}(\mathcal{T}^{\circ \ell})) \leq \binom{\ell + s - 1}{s - 1}.$$

■

Appendix B. Proof of Lemma 2

Lemma 1 *we have*

$$\|f(\mathcal{T}) - \mathcal{B}_{k,\ell}\|_\infty \leq \frac{\max_{|\eta| \leq 1} |f^{(\ell+1)}|(\eta)}{(\ell+1)!} \|(\mathcal{T} - \mathcal{O})^{\circ \ell + 1}\|_\infty \leq \left(\frac{k}{sm} \right)^{-(\ell+1)}.$$

Note that the covering number of s-dimensional bounded set $\mathcal{X}$ is $\mathcal{N}(1/k, \mathcal{X}, \|\cdot\|_\infty) \le k^s$. Let $\mathcal{E}_k$ denote the corresponding covering set. Then, $\mathcal{E}_k$ satisfies

1. $|\mathcal{E}_k| \le k^s$;
2. For every $\Delta \in \mathcal{E}_k$, we have
$$\max_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \Delta} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty \lesssim \frac{1}{k}.$$
3. $\Delta \cap \Delta' = \emptyset$ for all $\Delta \neq \Delta' \in \mathcal{E}_k$.

We label the center of the covering set by $\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{k^s}\} \subset \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}$. We use $z\colon [d] \to [k^s]$ to denote the membership of rows of $\boldsymbol{X}$

$$z(i) = \arg\min_{j \in [k^s]} \|\boldsymbol{x}_i - \boldsymbol{o}_j\|_\infty, \quad \text{such that} \quad \|\boldsymbol{x}_i - \boldsymbol{o}_{z(i)}\|_\infty \le \frac{1}{k}.$$

We define a block matrix $\mathcal{O} \in (\mathbb{R}^d)^{\otimes m}$ with entries

$$\mathcal{O}(i_1, \ldots, i_m) = \mathcal{C} \times_1 \boldsymbol{o}_{z(i_1)} \times_2 \cdots \times_m \boldsymbol{o}_{z(i_m)}$$

Therefore

$$\|\mathcal{T} - \mathcal{O}\|_\infty \le \max_{(i_1, \ldots, i_m) \in [d]^m} |\mathcal{C} \times_1 \boldsymbol{x}_{i_1} \times_2 \cdots \times_m \boldsymbol{x}_{i_m} - \mathcal{C} \times_1 \boldsymbol{o}_{z(i_1)} \times_2 \cdots \times_m \boldsymbol{o}_{z(i_m)}| \lesssim mk^{-1}s^{1/2}.$$

Define

$$\mathcal{B}_{k,\ell} := f^{(1)}(\mathcal{O}) \circ (\mathcal{T} - \mathcal{O}) + \frac{f^{(2)}(\mathcal{O})}{2} \circ (\mathcal{T} - \mathcal{O})^{\circ 2} + \cdots \frac{f^{(\ell)}(\mathcal{O})}{\ell!} \circ (\mathcal{T} - \mathcal{O})^{\circ \ell}.$$

The membership partition $[d]^m$ into k^{sm} blocks. We use $[d]^m = \cup_{n=1}^{k^{sm}} \Delta_n$ to denote these blocks. Within each block, the tensor $\mathcal{O}$ takes the same value,

$$\mathcal{B}_{k,\ell}(\omega) = \underbrace{(a_{0,\Delta} + a_{1,\Delta}\mathcal{T} + a_{2,\Delta}\mathcal{T}^{\circ 2} + \cdots a_{\ell,\Delta}\mathcal{T}^{\circ \ell})}_{:=\mathcal{I}} \mathbb{1}\{\omega \in \Delta\}. \tag{7}$$

Because there are k^s blocks along mode 1, we conclude that

$$\text{Trank}(\mathcal{B}_{k,\ell}) \le k^s \sum_{n=1}^{\ell} \binom{n+s-1}{s-1} \le k^s(\ell + s - 1)^s.$$

Appendix C. My Proof of Proposition 1

Lemma 2 (Polynomial and Unfolding) *Consider a function $f\colon (\mathbb{R}^s)^{\otimes m} \to \mathbb{R}$. Suppose that, under $(\mathbb{R}^s)^{\otimes m} \cong \mathbb{R}^{sm}$, the function f is a degree-ℓ polynomial of (sm)-variables; i.e.*

$$g(\boldsymbol{x}_1, \dots, \boldsymbol{x}_m) = \sum_{\kappa_1 + \cdots + \kappa_{sm} \le \ell} c(\kappa_1, \dots, \kappa_{sm}) x_{11}^{\kappa_1} \cdots x_{sm}^{\kappa_{sm}},$$

where x_{ij} denotes the i-th element in $\boldsymbol{x}_j$, for $(i,j) \in [s] \times [m]$. Then the dimensional-d order-m tensor $\boldsymbol{G}$ generated by f and $(\boldsymbol{a}_i)_{i\in[d]}$ admits

$$\boldsymbol{G}(i_1, \dots, i_m) = g(\boldsymbol{a}_{i_1}, \dots, \boldsymbol{a}_{i_m}), \quad \textit{for all } (i_1, \dots, i_m) \in [d]^m.$$

Then

$$\text{Trank}(\boldsymbol{G}) \le \sum_{i=0}^{\ell} \binom{i+s-1}{s-1}.$$

Proof

$$\text{Unfold}(\boldsymbol{G}) = \begin{bmatrix} \Phi(\boldsymbol{a}_1) \\ \vdots \\ \Phi(\boldsymbol{a}_d) \end{bmatrix} \boldsymbol{B}^T = \begin{bmatrix} 1 & a_{11} & \cdots & a_{s1} & a_{11}^2 & a_{11}a_{s1} & \cdots & a_{11}a_{21}a_{s1}^{\ell-2}\cdots & a_{s1}^{\ell} \\ \vdots & & & & & & & & \\ 1 & a_{1d} & \cdots & a_{sd} & a_{1d}^2 & a_{1d}a_{1d} & \cdots & a_{1d}a_{2d}a_{sd}^{\ell-2}\cdots & a_{sd}^{\ell} \end{bmatrix} \boldsymbol{B}^T$$

■

Lemma 3 (Uniform approximation) *Consider $\Theta \in \mathcal{P}(d, s, L)$. There exists a sequence of tensor $\{\mathcal{A}_{\ell,k}\colon (k,\ell) \in [d] \times \mathbb{N}\}$ such that, for every number of pieces $k \in [d]$ and degree $\ell \in \mathbb{N}$, we have*

$$\|\Theta - \mathcal{A}_{k,\ell}\|_\infty \le L k^{-\ell} \ell^{ms}, \quad \textit{and} \quad \text{Trank}(\mathcal{A}_{k,\ell}) \le k^s \ell^s.$$

Proof Based on the definition of infinitely smooth function

$$\begin{aligned} |\Theta - \mathcal{A}_{k,\ell}| &\le |f(\boldsymbol{a}_{i_1}, \dots, \boldsymbol{a}_{i_m}) - g(\boldsymbol{a}_{i_1}, \dots, \boldsymbol{a}_{i_m})| && (8) \\ &\le L \left(\frac{k}{a}\right)^{-\ell} \binom{\ell + ms - 1}{ms - 1} && (9) \\ &\le L \left(\frac{k}{a}\right)^{-\ell} \ell^{ms} && (10) \end{aligned}$$

■

Corollary 3 (Nice tensor model is of $\log d$ rank) *For any fixed $\varepsilon > 0$, we have*

$$\text{Trank}_\varepsilon(\Theta) := \min\{\text{Trank}(\mathcal{A})\colon \|\mathcal{A} - \Theta\|_\infty \le \varepsilon\} \lesssim \log d.$$

Where does d come from??

Theorem C.1 (Dimension-free approximation) *Let $\Theta \in \mathcal{P}(d, s, L)$. For every rank $r \in \mathbb{N}_+$,*

$$\frac{1}{d^m}\|\Theta,\ \mathit{Proj}_r(\Theta)\|_F^2 \lesssim L^2 \exp(-c_0 r^{1/s}).$$

Proof Let $k = \Omega(2a)$ and $\ell = r^{1/s}k^{-1}$. Then

$$\|\Theta - \mathcal{A}_{k,\ell}\|_\infty \leq L2^{-\ell}\ell^{ms} \lesssim L2^{-\ell} \asymp L\exp(-r^{1/s})$$

■

Appendix D. My Proof of Theorem 2

This is a complete version of a proof sketched in the main text.

Lemma 4 (Tensor polynomial approximation) *Suppose that 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 any $k \geq d$, there exists a k-membership function $z\colon [d] \to [k]$ and a polynomial function such that*

$$\Theta(i_1, \ldots, i_K) - P_{\lfloor\alpha\rfloor}(\omega - \mathcal{S}(z(i_1), \ldots, z(i_K))) \leq \frac{m^2}{k^\alpha}$$

,

$$\boldsymbol{M}(\boldsymbol{x} - \boldsymbol{x}_0) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 - x_0 & x_2 - x_0 & \cdots & x_d - x_0 \\ \vdots & \vdots & \vdots & \\ (x_1 - x_0)^{\lfloor\alpha\rfloor} & (x_2 - x_0)^{\lfloor\alpha\rfloor} & \cdots & (x_d - x_0)^{\lfloor\alpha\rfloor} \end{bmatrix}$$

Let $z\colon [d] \to [k]$ denote the clustering function. For notational convenience, we use the shorthand $\boldsymbol{z}(i_1, \ldots, i_m) := (z(i_1), \ldots, z(i_m))^T \in [k]^m$ to denote the block membership by applying z to each of the m modes. We make the convention that blockwise constant tensor is of degree 1 (not zero, for notational convenience).

1. blockwise degree-1 (constant) tensor

$$\mathscr{L}(0,k) = \left\{\mathcal{B} \in (\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(i_1,\ldots,i_m) = \mathcal{C}(z(i_1),\ldots,z(i_k)) \text{ for some tensor } \mathcal{C} \in (\mathbb{R}^k)^{\otimes m}\right\} \tag{11}$$

$$= \left\{\mathcal{B} \in (\mathbb{R}^d)^{\otimes m}\colon \Theta(\omega) = \sum_{\Delta\in[k]^m} c_\Delta \mathbb{1}\{\boldsymbol{z}(\omega) = \Delta\}\right\} \tag{12}$$

$$\cong \{\mathcal{C} \in \mathbb{R}^{k^m}, z\colon [d] \to [k]\} \tag{13}$$

   where, for every cluster $\Delta \in [k]^m$, the coefficient $c_\Delta \in \mathbb{R}$ represents the mean within the cluster. We use $\mathcal{C} = \{c_\Delta\colon \Delta \in [k]^m\}$ to collect all unknown coefficients.

2. blockwise degree-2 linear tensor

$$\mathscr{L}(1,k) = \left\{\mathcal{B} \in (\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(\omega) = \sum_{\Delta\in[k]^m} \left[c_\Delta + \langle \boldsymbol{\beta}_\Delta, \omega\rangle\right] \mathbb{1}\{\boldsymbol{z}(\omega) = \Delta\}\right\} \tag{14}$$

$$\cong \{\mathcal{C} \in \mathbb{R}^{(1+m)\times k^m}, z\colon [d] \to [k]\} \tag{15}$$

   where, for every cluster $\Delta \in [k]^m$, the coefficient $(c_\Delta, \boldsymbol{\beta}_\Delta) \in \mathbb{R} \times \mathbb{R}^d$ the mean and slope within the cluster. We use $\mathcal{C} = \{(c_\Delta, \boldsymbol{\beta}_\Delta)\colon \Delta \in [k]^m\} \cong \mathbb{R}^{(1+m)\times k^m}$ to collect all unknown coefficients.

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

$$\mathscr{L}(\ell+1,k) = \left\{\mathcal{B} \in (\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(\omega) = \sum_{\Delta\in[k]^m} \mathrm{Poly}_{\ell,\Delta}(\omega) \mathbb{1}\{\boldsymbol{z}(\omega) = \Delta\}\right\} \tag{16}$$

$$\cong \{\mathcal{C} \in \mathbb{R}^{(\ell+m)^\ell \times k^m}, z\colon [d] \to [k]\} \tag{17}$$

   where, for every cluster $\Delta \in [k]^m$, the coefficient $(c_\Delta, \boldsymbol{\beta}_\Delta) \in \mathbb{R} \times \mathbb{R}^d$ the mean and slope within the cluster. Note that the polynomial function $\mathrm{Poly}_{\ell,\Delta}(\cdot)$ has at most $(\ell+m)^\ell$ unknown coefficients. We use $\mathcal{C} \subset \mathbb{R}^{(\ell+m)^\ell \times k^m}$ to collect the unknown coefficients in the blockwise degree-$(\ell+1)$ polynomial tensor family.

Model:

$$\Theta(i_1,\ldots,i_m) = f\left(\frac{\pi(i_1)}{d},\ldots,\frac{\pi(i_m)}{d}\right), \quad \text{for all } i_1 < \cdots < i_d. \tag{18}$$

Theorem D.1 *(Blockwise polynomial tensor approximation) Suppose the function $f\colon [0,1]^m \to \mathbb{R}$ generating the signal tensor Θ as in model (18) 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*

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

We propose the least-square estimate

$$\hat{\Theta}^{\mathrm{LSE}} = \operatorname*{arg\,min}_{\Theta\in\mathscr{L}(\ell,k)} \|\mathcal{Y} - \Theta\|_F.$$

Rigorously, the least-square estimate $\hat{\Theta}^{\mathrm{LSE}}$ is a family of estimates with tuning parameters (ℓ, k); we suppress the dependence on (ℓ, k) when no confusion arises.

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

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

Theorem D.3 (Polynomial estimator) *Suppose that the signal tensor Θ is generated in model (18) with $f \in \mathcal{H}(\alpha) \cap \mathcal{M}(\beta)$. Let $\hat{\Theta}^{\mathrm{BC}}$ be the estimator in with degree $\ell^* = \lfloor \min(\alpha, \frac{m(m-1)}{2}) \rfloor$ and block size $k^* = \lfloor d^{\frac{m}{m+2\ell^*}} \rfloor$. Then the estimator $\hat{\Theta}^{BC}$ satisfies*

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

with very high probability.

Remark 1 *For order-3 tensor with sufficiently smooth function $\alpha \ge 3$, the optimal choice of degree and block sizes is $(\ell^*, k^*) = (3, d^{1/3})$.*

Theorem D.4 (Minimax lower bound) *For any given $\alpha \in (0,\infty)$, the problem of estimating α-smooth tensors obeys the minimax lower bound*

$$\mathbb{P}\left(\inf_{\hat{\Theta}} \max_{\Theta\in\mathcal{P}(\alpha)} \|\Theta - \hat{\Theta}\|_F^2 \ge d^{-\frac{2m\alpha}{m+2\alpha}} + d^{-(m-1)}\log d \right) > 0.8.$$