Smooth tensor estimation

Miaoyan Wang, May 18, 2020

1 Model

Key: Cartesian product of piecewise constant function representations. (has full generality...)

Let $\mathcal{Y} = [\![y_{i_1,\ldots,i_K}]\!] \in \{0,1\}^{d_1\times\cdots\times d_K}$ be an order-K, $(d_1,\ldots,d_K)$-dimensional binary tensor. Let $\boldsymbol{\xi}^{(k)} = (\xi_1^{(k)},\ldots,\xi_d^{(k)}) \in [0,1]^d$ be random vectors following (unknown) distributions $\mathbb{P}^{(k)}$ for all $k\in[K]$, and $\boldsymbol{\xi}^{(k)}$ and $\boldsymbol{\xi}^{(k')}$ are mutually independent for $k\neq k'\in[K]$. Assume that, conditional on $\{\boldsymbol{\xi}^{(k)}\}$, the entries of $\mathcal{Y}$ are independent sub-Gaussian distributed:

$$\mathbb{E}\left(y_{i_1,\ldots,i_K}|\boldsymbol{\xi}\right) = f\left(\xi_{1,i_1},\ldots,\xi_{K,i_K}\right), \quad \text{for all } (i_1,\ldots,i_K)\in[d]\times\cdots\times[d],$$

where $f\colon [0,1]^K \mapsto [0,1]$ is an unknown multivariate function belonging to a function class $f\in\mathcal{F}_\alpha(L)$. Specifically, the function class is defined as

$$\mathcal{F}_\alpha(L) = \{f\colon \mathrm{Im}(f)\in[0,1] \text{ and } \|f\|_{\mathcal{H}_\alpha}\leq L\},$$

where $\alpha\in(0,1]$ is the smoothness parameter and $L>0$ is the Hölder norm bound for the functions in the class.

Recall that the function Hölder norm $\|f\|_{\mathcal{H}_\alpha}$ is defined as

$$\|f\|_{\mathcal{H}_\alpha} \stackrel{\text{def}}{=} \max_{|\omega|\leq\lfloor\alpha\rfloor}\sup_{\boldsymbol{x}\in\mathcal{D}}|\nabla_\omega f(\boldsymbol{x})| + \max_{|\omega|=\lfloor\alpha\rfloor}\sup_{\boldsymbol{x}\neq\boldsymbol{x}'\in\mathcal{D}}\frac{\nabla_\omega|f(\boldsymbol{x})-\nabla_\omega f(\boldsymbol{x}')|}{\|\boldsymbol{x}-\boldsymbol{x}'\|_1^{\alpha-\lfloor\alpha\rfloor}},$$

where we have used the short-hand notion

$$\nabla_\omega f(\boldsymbol{x}) = \frac{\partial^{i_1+\cdots+i_K}}{\partial x_1^{i_1}\cdots\partial x_K^{i_K}} f(x_1,\ldots,x_k),$$

for multi-indices $\omega=(i_1,\ldots,i_K)$ with $|\omega| = i_1+\cdots+i_K$, and $\boldsymbol{x}=(x_1,\ldots,x_K)$ in the function domain.

2 Estimation

Define the objective function

$$F(\mathcal{C},\{\boldsymbol{M}_k\}) = \|\mathcal{Y}-\mathcal{C}\times_1\boldsymbol{M}_1\times\cdots\times_K\boldsymbol{M}_K\|_F^2.$$

Denote $\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times \cdots \times_K \boldsymbol{M}_K$ and $\boldsymbol{r} = (r_1, \ldots, r_K)$. Then the feasible domain is

$$\mathcal{P}(r) = \Big\{ \Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} \colon \Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times \cdots \times_K \boldsymbol{M}_K, \text{ where } \mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_K} \text{ and } \\ \boldsymbol{M}_k \in \{0,1\}^{d_k \times r_k} \text{ are membership matrices for all } k \in [K] \Big\}.$$

We propose constrained least-square estimator

$$\hat{\Theta}(\boldsymbol{r}, M) = \operatorname*{arg\,min}_{\Theta \in \mathcal{P}(\boldsymbol{r}), \|\Theta\|_\infty \leq M} F(\Theta \big| \mathcal{Y}).$$

The function estimator $\hat{f} \colon (0,1]^K \mapsto [0,1]$ is defined as follows:

$$\hat{f}(x_1, \ldots, x_K) \stackrel{\text{def}}{=} \hat{\Theta}(\lceil d_1 x_1 \rceil, \ldots, \lceil d_K x_K \rceil), \quad \text{for all } (x_1, \ldots, x_K) \in (0,1]^K.$$

We propose an adaptive smooth (?) estimation,

$$\hat{\Theta} = \operatorname*{arg\,min}_{\Theta \in \mathcal{P}(\boldsymbol{r}*)} F(\Theta), \quad \text{with} \quad \boldsymbol{r}^* = (r_1^*, \ldots, r_K^*), \\ \text{and} \quad r_k^* = \lceil d_k^{1/(\alpha \wedge 1 + 1)} \rceil \text{ for all } k \in [K].$$

Theorem 2.1. *Consider a function class $\mathcal{F}_\alpha(R)$ with $\alpha > 0$ and $M > 0$. We have*

$$\sup_{f \in \mathcal{F}_\alpha(R)} \sup_{\boldsymbol{\xi}^{(k)} \sim \mathbb{P}^{(k)}, k \in [K]} \frac{1}{d^K} \mathbb{E} \left(\|\hat{\Theta} - f(\xi_{i_1}^{(1)}, \ldots, \xi_{i_K}^{(K)})\|_F^2 \right) \leq C \left(d^{-K\alpha/(\alpha+1)} + \frac{\log d}{d^{K-1}} \right),$$

where the constant $C > 0$ depends only on L, and the expectation is taken jointly over $\mathcal{Y}$, $\{\boldsymbol{\xi}^{(k)}\}$ for all $k \in [K]$.

Phase transition at $\alpha = 1$ only for $K \geq 3$??

2.1 Non-parametric tensor model

Special cases: low-rank model, additive-multiplicative model (in the literature of non-parametric...)

Let $\mathcal{Y} = [\![Y_\omega]\!]$ be a binary tensor, where $\omega = (i_1, \ldots, i_K)$ is a K-tuple index. We propose the following conditionally-independent tensor model:

$$Y_\omega | \boldsymbol{\xi}_\omega \sim \text{Bernoulli}(\theta_\omega), \\ \theta_\omega = f(\xi_{i_1}^{(1)}, \ldots, \xi_{i_K}^{(K)}), \text{ for all } \omega = (i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K],$$

where $f \colon [0,1]^K \mapsto [0,1]$ is a multivariate function of interest. We use $\boldsymbol{\xi}_\omega \equiv (\xi_{i_1}^{(1)}, \ldots, \xi_{i_K}^{(K)})$ to denote the latent design variables at position $\omega = (i_1, \ldots, i_K)$. Furthermore, we assume that the collection of latent variables at the k-th coordinate, $(\xi_1^{(k)},\ldots,\xi_{d_K}^{(k)})$, follow a d_k-dimensional distribution $\mathbb{P}^{(k)}$, and the distributions $\mathbb{P}^{(k)}$ and $\mathbb{P}^{(k')}$ are mutually independent for $k\neq k'\in[K]$.

$$\begin{aligned}\boldsymbol{\xi}\colon [d_1]\times\cdots\times[d_K] &\mapsto [0,1]^K\\ \omega\equiv(i_1,\ldots,i_K) &\mapsto \boldsymbol{\xi}_\omega\equiv(\xi_{i_1}^{(1)},\ldots,\xi_{i_K}^{(K)})\sim\mathbb{P}^{(1)}\times\cdots\times\mathbb{P}^{(K)}.\end{aligned}$$

$$\begin{aligned}f\colon [0,1]^K &\mapsto [0,1]\\ \boldsymbol{\xi} &\mapsto f(\boldsymbol{\xi}).\end{aligned}$$

Lemma 1 (Connection between tensor block model and non-parametric tensor model)**.** *Let $f\in\mathcal{F}_\alpha(L)$ be a target function and $\boldsymbol{\xi}\sim\mathbb{P}_{\boldsymbol{\xi}}$ be realized latent variables. Let $f(\boldsymbol{\xi}_\omega)\in\mathbb{R}$ denote the tensor entry indexed by $\omega\in[d_1]\times\cdots\times[d_K]$, and let $f(\boldsymbol{\xi})=[\![f(\boldsymbol{\xi}_\omega)]\!]\in\mathbb{R}^{d_1\times\cdots\times d_K}$ denote the non-parametric tensor parameter. Then for each $\boldsymbol{r}\in[d_1]\times\cdots\times[d_K]$, there exists a parameter $\Theta=[\![\theta_\omega]\!]\in\mathcal{P}(\boldsymbol{r})$ such that*

$$\textit{Loss}(f(\boldsymbol{\xi}),\ \Theta)\le L^2\left(\sum_k\frac{1}{r_k}\right)^{\alpha\wedge 1}.$$

Remark 1. The (deterministic) bound holds uniformly over $\mathbb{P}_{\boldsymbol{\xi}}$ and $\mathcal{F}_\alpha(L)$.

Proof. The proof is constructive. We partition the interval $[0,1]$ into r_k equal-sized intervals at each of the K modes. Denote the grid of intervals $\mathcal{I}_{s_1,\ldots,s_K}=\left(\frac{s_1-1}{r_1},\frac{s_1}{r_1}\right]\times\cdots\times\left(\frac{s_K-1}{r_K},\frac{s_K}{r_K}\right]$ for all $(s_1,\ldots,s_K)\in[r_1]\times\cdots\times[r_K]$. We use the notation $\boldsymbol{\xi}_{\omega'}\sim\boldsymbol{\xi}_\omega$ if and only if $\boldsymbol{\xi}_{\omega'}$ and $\boldsymbol{\xi}_\omega$ belong to the same grid.

The parameter $\Theta=[\![\theta_\omega]\!]$ is constructed as follows. For each $\omega\in[d_1]\times\cdots\times[d_K]$, we set θ_ω to be the average of $f(\boldsymbol{\xi}_{\omega'})$ over the index set for which $\boldsymbol{\xi}_{\omega'}$ in the same grid as $\boldsymbol{\xi}_\omega$. Specifically, define

$$\theta_\omega=\frac{1}{c_\omega}\sum_{\omega'}f(\boldsymbol{\xi}_{\omega'})\mathbb{1}\{\boldsymbol{\xi}_{\omega'}\sim\boldsymbol{\xi}_\omega\},$$

where $c_\omega=\sum_{\omega'}\mathbb{1}\{\boldsymbol{\xi}_{\omega'}\sim\boldsymbol{\xi}_\omega\}$ is the number of tensor entries in the index set $\{\omega'\colon\boldsymbol{\xi}_{\omega'}\sim\boldsymbol{\xi}_\omega\}$. The construction implies that $\Theta=[\![\theta_\omega]\!]$ takes constant value in the index set $\{\omega\colon\boldsymbol{\xi}_\omega\in\mathcal{I}_{s_1,\ldots,s_K}\}$ for any given $(s_1,\ldots,s_K)\in[r_1]\times\cdots\times[r_K]$. Therefore, Θ is a block tensor with at most r blocks on each of the K modes; that is, $\Theta\in\mathcal{P}(\boldsymbol{r})$.

We will show that the defined Θ is close to $f(\boldsymbol{\xi}_\omega)$ in the square distance. Specifically,

$$|f(\boldsymbol{\xi}_\omega)-\theta_\omega|=\left|f(\boldsymbol{\xi}_\omega)-\frac{1}{c_\omega}\sum_{\omega'}f(\boldsymbol{\xi}_{\omega'})\mathbb{1}\{\boldsymbol{\xi}_{\omega'}\sim\boldsymbol{\xi}_\omega\}\right|$$

$$
\begin{aligned}
&\le \frac{1}{c_\omega} \sum_{\{\omega' : \boldsymbol{\xi}_{\omega'} \sim \boldsymbol{\xi}_\omega\}} |f(\boldsymbol{\xi}_\omega) - f(\boldsymbol{\xi}_{\omega'})| \\
&\le \frac{1}{c_\omega} \sum_{\{\omega' : \boldsymbol{\xi}_{\omega'} \sim \boldsymbol{\xi}_\omega\}} L \left\| \boldsymbol{\xi}_\omega - \boldsymbol{\xi}_{\omega'} \right\|^{\alpha \wedge 1} \\
&\le \frac{1}{c_\omega} \sum_{\{\omega' : \boldsymbol{\xi}_{\omega'} \sim \boldsymbol{\xi}_\omega\}} L \left(\frac{K}{r} \right)^{\alpha \wedge 1} \\
&\le LKr^{-(\alpha \wedge 1)},
\end{aligned}
$$

where the third line comes form the Hölder condition for f, the fourth line comes from the fact that $\|\boldsymbol{\xi}_\omega - \boldsymbol{\xi}_{\omega'}\| \le \frac{K}{r}$ for $\boldsymbol{\xi}_\omega \sim \boldsymbol{\xi}_{\omega'}$. Summing over all $\omega \in [d_1] \times \cdots \times [d_K]$ gives the conclusion

$$
\frac{1}{\prod_k d_k} \sum_\omega (f(\boldsymbol{\xi}_\omega) - \theta_\omega)^2 \le L^2 K^2 r^{-2(\alpha \wedge 1)}.
$$

□

Theorem 2.2 (MSE for tensor block model)**.** *Let $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$ be a binary tensor generated from tensor block model, i.e.*

$$
\mathcal{Y} | \Theta \sim \textit{Bernoulli}(\Theta), \quad \textit{where } \Theta \in \mathcal{P}(\boldsymbol{r}).
$$

Then for constant $C' > 0$, there exists a constant $C > 0$ such that

$$
\frac{1}{\prod_k d_k} \|\Theta - \hat{\Theta}\|_F^2 \le C \left(\prod_k \frac{r_k}{d_k} + \frac{\sum_k d_k \log r_k}{\prod_k d_k} \right),
$$

with probability at least $1 - \exp(-C' \sum_k d_k \log r_k)$, uniformly over $\Theta \in \mathcal{P}(\boldsymbol{r})$.

Suppose $r \asymp d^\delta$ for some $\delta \in [0,1]$.

$$
\text{MSE}(\Theta, \hat{\Theta}) \asymp \begin{cases} d^{-K}, & \delta = 0 \text{ and } r = 1, \\ d^{-(K-1)}, & \delta = 0 \text{ and } r \ge 2, \\ d^{-(K-1)} \log d, & \delta \in \left(0, \frac{1}{K}\right], \\ d^{-K(1-\delta)}, & \delta \in \left(\frac{1}{K}, 1\right]. \end{cases}
$$

Theorem 2.3. *Assume $r_k = O(r)$ and $d_k = O(d)$. Then*

$$
\textit{MSE}(f(\boldsymbol{\xi}), \hat{\Theta}) \le r^K + d \log r + L^2 K^2 r^{-2(\alpha \wedge 1)}.
$$

Proof. Taking $r = d^{\delta}$ where $\delta = \frac{K}{2(\alpha \wedge 1)+K}$ gives

$$\text{MSE}(f(\boldsymbol{\xi}), \hat{\Theta}) \leq \begin{cases} d^{-2\alpha/(2\alpha+1)} + o(1), & K = 1 \\ d^{-2\alpha/(\alpha+1)} + \frac{\log d}{d}, & K = 2 \\ d^{-2\alpha K/(2\alpha+K)} + d^{-2K/(K+2)}, & K \geq 3, \end{cases}$$

Only non-parametric rate appear for $K \geq 3$. $\square$

Remark 2. The distribution of $\mathbb{P}_{\boldsymbol{\xi}}$ is required to cover $[0,1]$. – in order to prove the lower bound..

We restrict to i.i.d. uniform distribution over $[0,1]$. (Erdos Roni)

Theorem 2.4. *There exists a constant $C > 0$ only depending on L, α, such that*

$$\inf_{\hat{\Theta}} \sup_{f \in \mathcal{F}_{\alpha}(L)} \sup_{\mathbb{P}_{\boldsymbol{\xi}} \in \mathcal{P}} \mathbb{E}\{MSE(\hat{\Theta}, f(\boldsymbol{\xi}))\} \begin{cases} d^{-2\alpha K/(2\alpha+K)}, & 0 < \alpha < 1, \\ d^{-1} \log d, & \alpha \geq 1 \text{ and } K = 2, \\ d^{-2K/(K+2)}, & \alpha \geq 1 \text{ and } K \geq 3. \end{cases}$$

3 Assumptions on function families

We consider two families of functions.

Definition 1 ($\mathcal{F}(R)$, piecewise constant functions with at most r marginal pieces)**.** A function $f\colon [0,1]^K \mapsto [0,1]$ is called a multivariate R-step function, if there exists a set of mappings $\{\phi_k\colon [0,1] \mapsto [R]\}_{k \in [K]}$ and an order-K tensor $\mathcal{C} \in \mathbb{R}^{R \times \cdots \times R}$ such that

$$f(x_1, \ldots, x_K) = \mathcal{C}(\phi_1(x_1), \ldots, \phi_K(x_K)), \quad \text{for all } (x_1, \ldots, x_K) \in [0,1]^K.$$

Denote $\boldsymbol{x} = (x_1, \ldots, x_K)^T$ and $\phi(\boldsymbol{x}) = (\phi_1(x_1), \ldots, \phi_K(x_K))^T$. Then f can be equivalently written as

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{r} \in [R]^K} \mathcal{C}(\boldsymbol{r}) \mathbb{1}\{\phi(\boldsymbol{x}) = \boldsymbol{r}\}, \text{ for all } \boldsymbol{x} \in [0,1]^K.$$

Remark 3. The number of constant pieces need not to be equal along each of the K modes. We use R to denote the upper bound for the number of constant pieces over the K modes.

Definition 2 ($\mathcal{F}(\alpha, L)$, Hölder smooth functions)**.** Let $\alpha \in (0,1]$ and $L > 0$. A function $f\colon [0,1]^K \mapsto [0,1]$ is called an α-Hölder smooth function if

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq L \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_1^{\alpha}, \quad \text{for all } \boldsymbol{x}, \boldsymbol{x}' \in [0,1]^K.$$

The constant $\alpha \in (0,1]$ is called the Hölder smoothness parameter and $L > 0$ is the Hölder constant.

Definition 3 (Measure-preserving bijection). Let $\tau\colon \mathcal{X} \mapsto \mathcal{X}$ be a bijection and U be a random variable taking values in $\mathcal{X}$. Then, τ is called a measure-preserving bijection with respect to U if for all $A \subset \mathcal{X}$.

$$\mathbb{P}(U \in A) = \mathbb{P}(\tau(U) \in A).$$

In particular, $\tau(U)$ and U are identically distributed.

Equivalent class ??

$$\{f'\colon f'(\boldsymbol{\xi}') \sim f(\boldsymbol{\xi}) \text{ whenever } \boldsymbol{\xi}' \sim \boldsymbol{\xi}\}$$

Definition 4 (Weakly isomorphism). Two functions $f, f'\colon [0,1]^K \mapsto [0,1]$ are called weakly isomorphic if there exists a set of measure-preserving bijections $\{\tau_k\colon [0,1] \mapsto [0,1]\}_{k\in[K]}$ such that

$$f(\xi_{1,i_1},\ldots,\xi_{K,i_K}) \stackrel{a.s.}{=} f'(\tau_1(\xi_{1,i_1}),\ldots,\tau_K(\xi_{K,i_K})), \quad \text{for all } (i_1,\ldots,i_K) \in [d_1]\times\cdots\times[d_K],$$

where $\{\xi_{k,i_k}\}$ are i.i.d. $U[0,1]$ for all $i_k \in [d_k]$ and all $k \in [K]$. We write $f \sim f'$ to denote weakly isomorphism. The weakly isomorphism relationship defines a quotient space in $\mathcal{F}(R)$ or $\mathcal{F}(\alpha, L)$.

Definition 5 (Index permutation). Two tensors $\Theta, \Theta' \in \mathbb{R}^{d_1\times\cdots\times d_K}$ are called equivalent if there exist a set of index permutations $\{\sigma_k : [d_k] \mapsto [d_k]\}_{k\in[K]}$, such that

$$\Theta(i_1,\ldots,i_K) = \Theta'(\sigma_1(i_1),\ldots,\sigma_K(i_K)), \quad \text{for all } (i_1,\ldots,i_K) \in [d_1]\times\cdots\times[d_K].$$

Two forms of loss are considered.

Definition 6 (Integrated loss). Let $f\colon [0,1]^K \mapsto [0,1]$ be the function of interest. We define the integrated loss

$$\text{Loss}(f,\hat{f}) \stackrel{\text{def}}{=} \inf_{f_{\text{iso}}\sim f} \int_{\boldsymbol{x}\in[0,1]^K} \left|f'(\boldsymbol{x}) - f_{\text{iso}}(\boldsymbol{x})\right|^2 d\boldsymbol{x},$$

where $f_{\text{iso}} \sim f$ denotes all (?need to be in the specified function space?) functions that are isomorphic with f.

Definition 7 (Discrete loss). Let $\Theta, \hat{\Theta} \in \mathbb{R}^{d_1\times\cdots\times d_K}$ be two tensors. We define the discrete loss as

$$\text{Loss}(\Theta,\hat{\Theta}) \stackrel{\text{def}}{=} \frac{1}{\prod_k d_K}\|\Theta - \hat{\Theta}\|_F^2.$$

Remark 4. We use the notion $\text{Loss}(\cdot,\cdot)$ to denote either the discrete loss (for tensors) or the integrated loss (for functions). The meaning should be clear given the contexts.

Definition 8 (Operations between f and Θ). Let $f\colon [0,1]^K \mapsto [0,1]$ be a K-variate function. Then the f-induced probability tensor Θ is defined as

$$\Theta(i_1,\ldots,i_K) = f(\xi_{1,i_1},\ldots,\xi_{K,i_K}), \quad \text{for all } (i_1,\ldots,i_K) \in [d_1]\times\cdots\times[d_K],$$

where $\{\xi_{k,i_k}\}$ are i.i.d. Uniform[0,1] for all $i_k \in [d_k]$ and all $k \in [K]$.

Conversely, let $\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ be an order-K tensor. Then the Θ-induced function is defined as

$$\begin{aligned} f(x_1, \ldots, x_K) &= \Theta(\lceil d_1 x_1 \rceil, \ldots, \lceil d_K x_K \rceil) \\ &= \sum_{(i_1,\ldots,i_K)} \Theta(i_1, \ldots, i_K) \mathbb{1}\{(i_k - 1) < x_k d_k \leq i_k, \text{ for all } k \in [K]\}, \end{aligned}$$

for all $(x_1, \ldots, x_K) \in [0,1]^K$.

Remark 5. The two operations are not inverse. $f \Rightarrow \Theta(f, \boldsymbol{\xi}) \Rightarrow f' \stackrel{\text{def}}{=} f(\Theta(f, \boldsymbol{\xi}))$, but $f \neq f'$ (not even isomorphic). Similarly, $\Theta \Rightarrow f(\Theta, \boldsymbol{\xi}) \Rightarrow \Theta' \stackrel{\text{ref}}{=} \Theta(f(\Theta, \boldsymbol{\xi}))$, but $\Theta' \neq \Theta$. The inverse relationship holds only if $\xi_{i_k,k} = \frac{i_k}{d_k}$ for all $i_k \in [d_k]$ and all $k \in [K]$.

Proposition 1 (Connection between f and Θ)**.** The following properties hold:

1. [From functions to tensors] Let Θ denote the induced tensor from f. Then, $f \in \mathcal{F}(R) \Rightarrow \Theta \in \mathcal{P}(R, 1)$. Furthermore, $\Theta' \sim \Theta \Rightarrow$ there exists f' such that Θ' is the induced tensor from f'.

2. [From tensors to functions] Let f denote the induced function from Θ. Then $\Theta \in \mathcal{P}(R, 1) \Rightarrow f \in \mathcal{F}(R)$.

3. [From tensor pairs to function pairs] Let $f,\ f'$ denote the induced functions from Θ and Θ', respectively. Then, $f \sim f' \Leftrightarrow \Theta \sim \Theta'$. Furthermore,

$$\text{Loss}(f,\ f') \leq \text{Loss}(\Theta,\ \Theta').$$

Problem 1 (Non-parametric estimation with unobserved designs)**.** Let $f\colon [0,1]^K \mapsto [-1,1]$ be a target function of interest. For each $k \in [K]$, we draw a random i.i.d. sample of d points $\{x_i^{(k)}\}_{i\in[d]}$ uniformly from $[0,1]$. Write $\omega = (i_1, \ldots, i_K) \in [d]^K$, $\boldsymbol{x}_\omega = (x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)})^T \in \mathbb{R}^K$, and $f_\omega = f(\boldsymbol{x}_\omega) \in \mathbb{R}$. The goal is to estimation f given the set $\{(\omega, f_\omega)\}$ but without knowing $\{\boldsymbol{x}_\omega\}$!

We estimate f by using the null design; that is, we partition $[0,1]^K$ into d^K equal-sized grids. Specifically, we assume $\check{x}_i^{(k)} = \frac{i}{d}$ for all $i \in [d]$ and $k \in [K]$, and write $\check{\boldsymbol{x}}_\omega = (\check{x}_{i_1}^{(1)}, \ldots, \check{x}_{i_K}^{(K)})^T \in \mathbb{R}^K$. Then the estimation is based on $\{(\check{\boldsymbol{x}}_\omega, f_\omega)\}$.

Q1: existing results for known $\{\boldsymbol{x}_\omega\}$? Q2: quantify the difference between $\{\boldsymbol{x}_\omega\}$ vs. $\{\check{\boldsymbol{x}}_\omega\}$.

The following results give the estimation error when f belongs to $\mathcal{F}(R, 1)$ or $\mathcal{F}(\alpha, L)$.

Proposition 2 (Agnostic error)**.** Consider the chain $f \Rightarrow \Theta(f, \boldsymbol{\xi}) \Rightarrow \hat{f} \stackrel{\text{def}}{=} f(\Theta(f, \boldsymbol{\xi}))$.

- If $f \in \mathcal{F}(R, 1)$, then
$$\mathbb{E}\left[\text{Loss}(f,\ \hat{f})\right] \leq \frac{CKR}{\sqrt{d}},$$
where the expectation is over $\boldsymbol{\xi} \sim \mathbb{P}_{\boldsymbol{\xi}}$.

- If $f \in \mathcal{F}(\alpha, L)$, then

$$\mathbb{E}\left[\text{Loss}(f,\ \hat{f})\right] \le \frac{CK^2L^2}{d^\alpha},$$

where the expectation is over $\boldsymbol{\xi} \sim \mathbb{P}_{\boldsymbol{\xi}}$.

Remark 6. Does it matter how to define $\hat{f}$? say kernel, or smooth version?

Proof. Note that

$$\Theta(i_1, \ldots, i_K) = f(\xi_{i_1}^{(1)}, \ldots, \xi_{i_K}^{(K)}), \quad \text{for all } (i_1, \ldots, i_K) \in [d]^K.$$

For each $k \in [K]$, we sort $\{\xi_i^{(k)}\}_{i \in [d]}$ in an increasing order and write $0 \le \xi_{(1)}^{(k)} \le \cdots \le \xi_{(d)}^{(k)} \le 1$. Define

$$\Theta^*(i_1, \ldots, i_K) = f(\xi_{(i_1)}^{(1)}, \ldots, \xi_{(i_K)}^{(K)}), \quad \text{for all } (i_1, \ldots, i_K) \in [d]^K.$$

Since $\Theta^* \sim \Theta$, they induce weakly isomorphic functions. Now consider the function induced by Θ^*.

$$\begin{aligned}
\hat{f}(\boldsymbol{x}) &= \sum_{(i_1, \ldots, i_K)} \Theta^*(i_1, \ldots, i_K) \mathbb{1}\left\{x_k \in \left(\frac{i_k - 1}{d},\ \frac{i_k}{d}\right]\right\} \\
&= \sum_{(i_1, \ldots, i_K)} f(\xi_{(i_1)}^{(1)}, \ldots, \xi_{(i_K)}^{(K)}) \mathbb{1}\left\{x_k \in \left(\frac{i_k - 1}{d},\ \frac{i_k}{d}\right],\ \text{for all } k \in [K]\right\}.
\end{aligned}$$

We aim to evaluate the integral over the grid $I_i = (\frac{i-1}{d}, \frac{i}{d}]$ for $i \in [d]$.

$$\text{Loss}(f,\ \hat{f}) = \sum_{(i_1, \ldots, i_K)} \left[\int_{\boldsymbol{x} \in I_{i_1} \times \cdots \times I_{i_K}} |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})|^2 d\boldsymbol{x}\right] \tag{1}$$

We now evaluate the integral over the region $I_\omega = I_{i_1} \times \cdots \times I_{i_K}$. Define $\check{\boldsymbol{x}} = (\frac{i_1-1}{d}, \frac{i_1}{d}] \times \cdots (\frac{i_K-1}{d}, \frac{i_K-1}{d}] \in I_\omega$. For any $\boldsymbol{x} \in I_\omega$, we have

$$|f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})| \le |f(\boldsymbol{x}) - f(\check{\boldsymbol{x}})| + \left|\hat{f}(\boldsymbol{x}) - f(\check{\boldsymbol{x}})\right|.$$

Note that both $\check{\boldsymbol{x}}, \boldsymbol{x} \in I_\omega$

$$|f(\boldsymbol{x}) - f(\check{\boldsymbol{x}})| \le L|\boldsymbol{x} - \check{\boldsymbol{x}}|^\alpha \le LKd^{-\alpha}. \tag{2}$$

Furthermore,

$$\begin{aligned}
|\hat{f}(\boldsymbol{x}) - f(\check{\boldsymbol{x}})| &\le |f(\xi_{(i_1)}^{(1)}, \ldots, \xi_{(i_K)}^{(K)}) - f(\frac{i_1}{d}, \ldots, \frac{i_K}{d})| \\
&\le L \max_k \left|\xi_{(i_k)}^{(k)} - \frac{i_k}{d}\right|^\alpha.
\end{aligned} \tag{3}$$

Plugging (2) and (3) to (1), we obtain

$$\int_{\boldsymbol{x}\in[0,1]^K} |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})|^2 d\boldsymbol{x} \le L^2K^2 \left\{ d^{-\alpha} + \max_{k,i_k} \left| \xi_{(i_k)}^{(k)} - \frac{i_k}{d} \right|^{\alpha} \right\}^2$$
$$\le L^2K^2 \left\{ d^{-2\alpha} + \max_{k,i_k} \left| \xi_{(i_k)}^{(k)} - \frac{i_k}{d} \right|^{2\alpha} + 2d^{-\alpha} \max_{k} \left| \xi_{(i_k)}^{(k)} - \frac{i_k}{d} \right|^{\alpha} \right\}.$$

By Jensen's inequality, $f(x) = x^\alpha$ is concave for $\alpha \in [0,1)$, so henece $\mathbb{E}[f(x)] \le f(\mathbb{E}(x))$:

$$\mathbb{E} \left| \xi_{(i_k)}^{(k)} - \frac{i_k}{d} \right|^{2\alpha} \le \left[\mathrm{Var} \left(\xi_{(i_k)}^{(k)} \right) \right]^{\alpha} \le Cd^{-\alpha}.$$

The last line comes from the fact that the order statistics of the uniform distribution belongs to Beta distribution, $\xi_{(i)} \sim \mathrm{Beta}(i, d+1-i)$ for all $i \in [d]$. Then $\mathrm{Var}(\xi_{(i)})$ By $\mathbb{E}(X) \le \sqrt{\mathbb{E}(X^2)}$, we have

$$\mathbb{E} \left| \xi_{(i_k)}^{(k)} - \frac{i_k}{d} \right|^{\alpha} \le Cd^{-\alpha/2}.$$

Therefore,

$$\mathrm{Loss}(f, \hat{f}) \le L^2K^2 \left(d^{-2\alpha} + Cd^{-\alpha} + 2Cd^{-3\alpha/2} \right) \le CL^2K^2d^{-\alpha}.$$

$\square$

Proof. By definition, $f \in \mathcal{F}(R,1)$ implies there exist a set of mappings $\phi_k \colon [0,1] \mapsto [R]$ such that

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{r}\in[R]^K} \mathcal{C}(\boldsymbol{r}) \mathbb{1}\left\{ \phi_k(x_k) = r_k, \text{ for all } k \in [K] \right\}, \text{ for all } \boldsymbol{x} \in [0,1]^K.$$

Without loss of generality, assume $\phi_k^{-1}(r) = (\lambda_{r-1}^{(k)}, \lambda_r^{(k)}]$, where $0 = \lambda_1^{(k)} \le \cdots \lambda_2^{(k)} \le \cdots \le \lambda_{R-1}^{(k)} \le \lambda_R^{(k)} = 1$ are a sequence of cut-off points on $[0,1]$, and the interval length preserves the Lebesgue measure $|\{x_k \in \phi_k^{-1}(r)\}| = |\lambda_r^{(k)} - \lambda_{r-1}^{(k)}|$ for all $r \in [K]$. Under this assumption, f has an isomorphic form

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{r}\in[R]^K} \mathcal{C}(\boldsymbol{r}) \mathbb{1}\left\{ x_k \in (\lambda_{r_k-1}^{(k)},\ \lambda_{r_k}^{k}], \text{ for all } k \in [K] \right\}, \text{ for all } \boldsymbol{x} \in [0,1]^K. \tag{4}$$

Now consider $f \Rightarrow \Theta$. By definition, for all $(i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K]$,

$$\Theta(i_1, \ldots, i_K) = \sum_{\boldsymbol{r}\in[R]^K} \mathcal{C}(\boldsymbol{r}) \mathbb{1}\left\{ \xi_{i_k}^{(k)} \in (\lambda_{r_k-1}^{(k)},\ \lambda_{r_k}^{(k)}], \text{ for all } k \in [K] \right\},$$

where $\{\xi_{i_k}^{(k)}\}$ are i.i.d. Unif[0,1] for all $i_k \in [d]$ and $k \in [K]$.

From $\Theta \Rightarrow \hat{f}$, we have

$$\begin{aligned}\hat{f}(\boldsymbol{x}) &= \sum_{(i_1,\ldots,i_K)} \Theta(i_1,\ldots,i_K)\mathbb{1}\left\{x_k \in \left(\frac{i_k-1}{d},\ \frac{i_k}{d}\right],\ \text{for all } k\in[K]\right\} \\ &= \sum_{\boldsymbol{r}\in[R]^K} \mathcal{C}(\boldsymbol{r}) \sum_{(i_1,\ldots,i_K)} \mathbb{1}\left\{\xi_{i_k}^{(k)} \in \left(\lambda_{r_k-1}^{(k)},\ \lambda_{r_k}^{(k)}\right] \text{ and } x_k \in \left(\frac{i_k-1}{d},\ \frac{i_k}{d}\right],\ \text{for all } k\in[K]\right\} \\ &= \sum_{\boldsymbol{r}} \mathcal{C}(\boldsymbol{r}) \prod_{k\in[K]} \left(\sum_{i_k\in[d]} \mathbb{1}\left\{\xi_{i_k}^{(k)} \in \left(\lambda_{r_k-1}^{(k)},\ \lambda_{r_k}^{(k)}\right] \text{ and } x_k \in \left(\frac{i_k-1}{d},\ \frac{i_k}{d}\right]\right\}\right)\end{aligned}$$

For each $k\in[K]$, define the empirical cumulative proportion of categories among the set $\{\xi_1^{(k)},\ldots,\xi_d^{(k)}\}$

$$\hat{\lambda}_r^{(k)} = \frac{1}{d}\sum_{i\in[d]} \mathbb{1}\left\{\xi_i^{(k)} \leq \lambda_r^{(k)}\right\},\ \text{for all } r\in[R].$$

The function $\hat{f}$ can be equivalently written as follows:

$$\begin{aligned}\hat{f}(\boldsymbol{x}) &= \sum_{\boldsymbol{r}} \mathcal{C}(\boldsymbol{r}) \prod_{k\in[K]} \mathbb{1}\left\{x_k \in (\hat{\lambda}_{r_k-1}^{(k)},\ \hat{\lambda}_{r_k}^{(k)}]\right\} \\ &= \sum_{\boldsymbol{r}} \mathcal{C}(\boldsymbol{r}) \mathbb{1}\{x_k \in (\hat{\lambda}_{r_k-1}^{(k)},\ \hat{\lambda}_{r_k}^{(k)}] \text{ for all } k\in[K]\}.\end{aligned} \tag{5}$$

Applying Lemma 3 to (4) and (5), we conclude

$$\mathbb{E}[\text{loss}(f,\ f')] \leq \frac{8KR}{\sqrt{d}}.$$

□

We partition the set $[0,1]^K$ in two ways:

1. Partition based on f.

$$[0,1]^K = \cup\left\{\phi^{-1}(\boldsymbol{r})\colon \boldsymbol{r}\in[R]^K\right\},\ \text{where } \phi^{-1}(\boldsymbol{r}) \stackrel{\text{def}}{=} \{\boldsymbol{x}\in[0,1]^K\colon \phi(\boldsymbol{x}) = \boldsymbol{r}\}.$$

2. Partition based on $\hat{f}$.

$$[0,1]^K = \cup\left\{h^{-1}(\boldsymbol{r})\colon \boldsymbol{r}\in[R]^K\right\},\ \text{where } h^{-1}(\boldsymbol{r}) \stackrel{\text{def}}{=} \{\boldsymbol{x}\in[0,1]^K\colon h(\boldsymbol{x},\boldsymbol{r}) = 1\}.$$

We define a one-to-one mapping between $\phi^{-1}(\boldsymbol{r})$ and $h^{-1}(\boldsymbol{r})$. The mapping is measure-preserving in the sense that $|p_{\boldsymbol{r}}|$ (??)

We will evaluate the difference between $\phi^{-1}(\boldsymbol{r})$ and $h^{-1}(\boldsymbol{r})$. The $p_{\boldsymbol{r}} := |\phi^{-1}(\boldsymbol{r})| \in [0,1]$ denote the Lebesgue measure of the set $\phi^{-1}(\boldsymbol{r})$. Note that $h^{-1}(\boldsymbol{r})$ is a random set where the randomness comes from $\boldsymbol{\xi}_\omega$. In particular,

$$
\begin{aligned}
\hat{p}_{\boldsymbol{r}} \overset{\text{def}}{=} |h^{-1}(\boldsymbol{r})| &= |\{\boldsymbol{x}\colon \sum_\omega \mathbb{1}\{\phi(\boldsymbol{\xi}_\omega) = \boldsymbol{r}\}\mathbb{1}\{\boldsymbol{x} \in I_\omega\}\}| \\
&= \frac{1}{d^K} \sum_\omega \mathbb{1}\{\phi(\boldsymbol{\xi}_\omega) = \boldsymbol{r}\} \\
&= \frac{1}{d^K} \sum_{(i_1,\ldots,i_K)} \left(\prod_k \mathbb{1}\{\phi_k(\xi_{k,i_k}) = r_k\} \right) \\
&= \frac{1}{d^K} \prod_k \left(\sum_{i_k \in [d]} \mathbb{1}\{\phi_k(\xi_{k,i_k}) = r_k\} \right) \\
&= \frac{1}{d^K} \prod_k \mathrm{Bin}(d,\ \lambda_k(r_k))
\end{aligned}
$$

where the event $\mathbb{1}\{\phi_k(\xi_{k,i_k}) = r_k\}$ are i.i.d. Bernoulli random variables with success probability $\lambda_k(r_k) = |x_k \in [0,1]\colon \phi_k(x_k) = r_k| \in [0,1]$ for all $i_k \in [d]$ and $k \in [K]$. Therefore,

$$
\hat{p}_{\boldsymbol{r}} \sim \frac{1}{d^K} \prod_k \mathrm{Bin}(d, \lambda_k(r_k)) \quad \text{and} \quad p_{\boldsymbol{r}} = \prod_k \lambda_k(r_k).
$$

which implies

$$
\mathbb{E}\left(|\hat{p}_{\boldsymbol{r}} - p_{\boldsymbol{r}}|^2\right) \leq \frac{K p_{\boldsymbol{r}}}{d}.
$$

Now, we evaluate the loss:

$$
\int_{\boldsymbol{x} \in [0,1]^K} |f(\boldsymbol{x}) - f'(\tau(\boldsymbol{x}))|^2 d\boldsymbol{x} \leq \sum_{\boldsymbol{r} \in [R]^K} \int_{\boldsymbol{x} \in \phi^{-1}(\boldsymbol{r})} |f(\boldsymbol{x}) - f'(\tau(\boldsymbol{x}))|^2 d\boldsymbol{x}
$$

Proof.

$$
\begin{aligned}
\mathrm{Loss}(f,\ f') &= \inf_{f'_{\text{iso}} \sim f' \sim f} \int_{\boldsymbol{x} \in [0,1]^K} |f'_{\text{iso}}(\boldsymbol{x}) - f(\boldsymbol{x})|^2 d\boldsymbol{x} \\
&\leq \inf_{\tau\colon \text{ measure-preserving map}} \int_{\boldsymbol{x} \in [0,1]^K} |f'(\tau(\boldsymbol{x})) - f(\boldsymbol{x})|^2 d\boldsymbol{x}.
\end{aligned}
$$

By assumption, f is piecewise constant over the gird, $\mathcal{G} = \cup_i I_i$, where the intervals $I_i \subset [0,1]^K$ are disjoint for all $i \in [R^K]$. Note that

$$
\int_{\boldsymbol{x} \in I_i} |f(\boldsymbol{x}) - f'(\boldsymbol{x})|^2 d\boldsymbol{x} = 4 \int_{\boldsymbol{x} \in I_i} \mathbb{1}\left\{f(\boldsymbol{x}) \neq f'(\boldsymbol{x})\right\} d\boldsymbol{x} = 4\mu\{\boldsymbol{x} \in I_i\colon f(\boldsymbol{x}) \neq f'(\boldsymbol{x})\},
$$

where $\mu\{\cdot\}$ denotes the Lebesgue measure. We aim to find two isomorphisms to upper bound the Lebesgue measure. Specifically, let $c = f(\boldsymbol{x})$ denote the constant in the interval I_i. We want to find the region for which $\{\boldsymbol{x} \in I_i \colon f'(\tau(\boldsymbol{x})) \neq c\}$ where τ is a measure-preserving map. Recall the definition $f \Rightarrow \Theta(f, \boldsymbol{\xi}) \Rightarrow f'$.

First, we consider $f \Rightarrow \Theta(f, \boldsymbol{\xi})$. The random tensor $\Theta = \Theta(f, \boldsymbol{\xi})$ is expressed as

$$\Theta(i_1, \ldots, i_K) = f(\xi_{1,i_1}, \ldots, \xi_{K,i_K}), \text{ for all } [i_1, \ldots, i_K] \in [d_1] \times \cdots \times [d_K].$$

For each fixed $k \in [K]$, we sort the elements in $\{\xi_{k,i_k} \colon i_k \in [d_k]\}$ from smallest to largest. With a little abuse of notation, we denote $\xi_{k,1} \leq \cdots \leq \xi_{k,d_k}$ for all $k \in [K]$. The sorted tensor Θ^* is expressed as

$$\Theta^*(i_1, \ldots, i_K) = \Theta(\sigma_1(i_1), \ldots, \sigma_K(i_K)), \text{ for all } [i_1, \ldots, i_K] \in [d_1] \times \cdots \times [d_K],$$

where $\sigma_k \colon [d_k] \mapsto [d_k]$ denotes the permutation that sorts $\{\xi_{k,i_k}\}_{i_k \in [d_k]}$ in increasing order. Note that $\Theta^* \sim \Theta$. By property (i), there exists $f^* \sim f$ such that Θ^* is induced by f^*. Therefore, it suffices to evaluate the loss between functions f^* and f'. Furthermore, by property (ii), without loss of generality, we define f' the function induced by Θ^*. We aim to investigate the value $f'(\boldsymbol{x})$ for $\boldsymbol{x} \in I$.

Then, we consider $\Theta^* \Rightarrow f'$. By definition, for all $(x_1, \ldots, x_K) \in [0,1]^K$,

$$\begin{aligned}
f'(\boldsymbol{x}) &= \sum_{(i_1,\ldots,i_K)} \Theta^*(i_1, \ldots, i_K) \mathbb{1}\left\{\boldsymbol{x} \in \left(\frac{i_1 - 1}{d_1}, \frac{i_1}{d_1}\right] \times \cdots \times \left(\frac{i_K - 1}{d_K}, \frac{K}{d_K}\right]\right\}, \\
&= \sum_{(i_1,\ldots,i_K)} f(\xi_{1,i_1}, \ldots, \xi_{K,i_K}) \mathbb{1}\left\{\boldsymbol{x} \in \left(\frac{i_1 - 1}{d_1}, \frac{i_1}{d_1}\right] \times \cdots \times \left(\frac{i_K - 1}{d_K}, \frac{K}{d_K}\right]\right\} \\
&= \sum_{\boldsymbol{r}} C(\boldsymbol{r}) \mathbb{1}\{\phi(\boldsymbol{x}) = \boldsymbol{r}\}
\end{aligned}$$

Recall that $\Theta \in \mathcal{P}(R, 1)$.... Suppose $\xi_1 < \cdots < \xi_d$. $\xi_k \in R$, $\xi_{k+1} \in R$,... then $d_1 x_1$

The only places that f and f' differ are

$$|I - \hat{I}|, \quad \text{where } d_{\text{total}} \hat{I} \sim \text{Bernoulli}(d_{\text{total}}, I).$$

Therefore

$$\mathbb{E}(|I - \hat{I}|) \leq \sqrt{\frac{I}{d_{\text{total}}}}.$$

Summing over I_i over $i \in [R^K]$ and note that $\sum_i I_i = 1$, we have

$$\left|\{\boldsymbol{x} : f(\boldsymbol{x}) \neq f'(\boldsymbol{x})\}\right|_\lambda \leq \sum_{i \in [R^K]} \mathbb{E}(|I_i - \hat{I}_i|) \leq \sum_{i \in [R^K]} \frac{\sqrt{I_i}}{\sqrt{d^K}} \leq \left(\frac{R}{d}\right)^{K/2}.$$

$\square$

4 Main results

Theorem 4.1 (loss in $\mathcal{F}(R)$ family)**.** *Let $f \in \mathcal{F}(R)$ and $\mathcal{Y} \sim \mathbb{P}(f, \boldsymbol{\xi})$ generated from the tensor nonparametric model. Let $\hat{f}$ be the least-square estimator with pre-specified block size R. Then,*

$$\mathbb{E}\left[Loss(f, \hat{f})\right] \leq C\left(\frac{R^K}{d^K} + \frac{K \log R}{d^{K-1}}\right) + \frac{8KR}{\sqrt{d}},$$

where the expectation is taken jointly over $\mathcal{Y}$ and $\boldsymbol{\xi}$ (??) (uniformly over $f \in \mathcal{F}(R)$??). In particular, let $\Theta = f(\boldsymbol{\xi})$ and $\hat{\Theta} = \hat{f}(\boldsymbol{\xi})$. Then,

$$\mathbb{E}\left[Loss(\Theta, \hat{\Theta})\right] \leq C\left(\frac{R^K}{d^K} + \frac{\log R}{d^{K-1}}\right), \tag{6}$$

where again the expectation is taken jointly over $\mathcal{Y}$ and $\boldsymbol{\xi}$ (? or conditinoally).

Proof. We first proof (6). Recall that $\hat{\Theta} = \arg\min_{\Theta \in \mathcal{F}(R,M)} F(\Theta)$ and $\omega_{\Theta \in \mathcal{F}(R,M)} \nabla^2 F(\Theta) = \frac{1}{2}$. By Taylor expansion of $F(\Theta)$ around $\hat{\Theta}$, we have

$$\begin{aligned}
\|\hat{\Theta} - \Theta^{\text{true}}\|_F &\leq 2\left\langle \Theta^{\text{true}} - \mathcal{Y},\ \frac{\hat{\Theta} - \Theta^{\text{true}}}{\|\hat{\Theta} - \Theta^{\text{true}}\|_F} \right\rangle \\
&\leq 2 \max_{\Theta \in \mathcal{P}(2R,1)} \langle \mathcal{E},\ \Theta \rangle.
\end{aligned}$$

By union bound, for any $t > 0$

$$\begin{aligned}
\mathbb{P}\left(\max_{\Theta^{\text{true}} \in \mathcal{P}(R,M)} \|\hat{\Theta} - \Theta^{\text{true}}\|_F \geq t\right) &\leq \sup_{\boldsymbol{\xi}} \mathbb{P}\left(\max_{\Theta \in \mathcal{P}(2R,1)} \langle \mathcal{E},\ \Theta \rangle \geq \frac{t}{2} | \boldsymbol{\xi}\right) \\
&\leq \sup_{\boldsymbol{\xi}} C |\mathcal{P}(2R,1)| \exp\left(-\frac{t^2}{4\sigma^2}\right) \\
&\leq C \exp\left(-\frac{t^2}{\sigma^2} + Kd \log R + R^K\right).
\end{aligned}$$

We take $t = \sigma\sqrt{Kd \log R + R^K}$ and obtain that,

$$\max_{\Theta^{\text{true}} \in \mathcal{P}(R,M)} \|\hat{\Theta} - \Theta^{\text{true}}\|_F^2 \leq \sigma^2 (Kd \log R + R^K),$$

with probability at least $1 - \exp(-Kd \log R - R^K)$, (or equivalently, uniformly over $f \in \mathcal{F}(R, M)$ and $\boldsymbol{\xi} \in \mathbb{P}_{\boldsymbol{\xi}}$). Therefore,

$$\mathbb{E}(\text{Loss}(\hat{\Theta}, \Theta^{\text{true}})) \leq \sigma\left(\frac{r^K}{d^K} + \frac{K \log r}{d^{K-1}}\right),$$

where the expectation with respect to $\mathcal{Y}$ (and $\boldsymbol{\xi}$). $\square$

Lemma 2.

$$\textit{Loss}(f, \hat{f}) \leq \textit{Loss}(\Theta, \hat{\Theta}) + \frac{2KR}{\sqrt{d}}.$$

Proof. Define $f \Rightarrow \Theta(f, \boldsymbol{\xi}) \Rightarrow \check{f}$ (a random function measure-able w.r.t. $\boldsymbol{\xi}$).

$$\begin{aligned} \mathbb{E}\left[\text{Loss}(f, \hat{f})\right] &\leq \mathbb{E}\left[\text{Loss}(\tilde{f}, \hat{f})\right] + \mathbb{E}\left[\text{Loss}(\tilde{f}, f)\right] \\ &\leq \mathbb{E}\left[\text{Loss}(\Theta, \hat{\Theta})\right] + \mathbb{E}\left[\text{Loss}(\tilde{f}, f)\right] \\ &\leq \sigma^2\left(\frac{K\log R}{d^{K-1}} + \frac{R}{d^K}\right) + \frac{2KR}{\sqrt{d}}, \end{aligned}$$

where the expectation is over $\mathcal{E}$ and $\boldsymbol{\xi}$ jointly. □

Remark 7. Dense region: agonistic error dominates. Sparse region: estimation error dominates... intuition?

Problem 2 (Non-parametric estimation with unobserved designs)**.** Let $f\colon [0,1]^K \mapsto [-1,1]$ be a target function of interest. For each $k \in [K]$, we draw a random i.i.d. sample of d points $\{x_i^{(k)}\}_{i\in[d]}$ uniformly from $[0,1]$. Write $\omega = (i_1, \ldots, i_K) \in [d]^K$ and $\boldsymbol{x}_\omega = (x_{i_1}^{(1)}, \ldots, x_{i_K}^{(K)})^T \in \mathbb{R}^K$. The goal is to estimation f given the set $\{(\omega, f(\boldsymbol{x}_\omega))\}$ but without knowing $\{\boldsymbol{x}_\omega\}$!

The following lemma gives the estimation error for stepwise constant functions $f \in \mathcal{F}(R, 1)$.

Lemma 3 (Step function approximation with unobserved designs)**.** *Let $f\colon [0,1]^K \mapsto [-1,1]$ be an R-step function:*

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{r}\in[R]^K} c_{\boldsymbol{r}} \mathbb{1}\left\{\boldsymbol{x} \in \underbrace{(\lambda_{r_1-1}^{(1)},\ \lambda_{r_1}^{(1)}] \times \cdots \times (\lambda_{r_K-1}^{(K)},\ \lambda_{r_K}^{(K)}]}_{=:I_{\boldsymbol{r}}}\right\}, \quad \textit{for all } \boldsymbol{x} \in [0,1]^K,$$

where $\boldsymbol{r} = (r_1, \ldots, r_K) \in [R]^K$ denotes multi-index, $\{c_{\boldsymbol{r}} \in [-1,1]\}$ are a set of real numbers, and $0 = \lambda_1^{(k)} \leq \lambda_2^{(k)} \leq \cdots \leq \lambda_{R-1}^{(k)} \leq \lambda_R^{(k)} = 1$ are a sequence of cutoff points over $[0,1]$, for $k \in [K]$. Let $p_r^{(k)} = \lambda_r^{(k)} - \lambda_{r-1}^{(k)}$ denote the length of the r-th interval at the mode k, where $r \in [R]$ and $k \in [K]$.

For each $k \in [K]$, draw a random i.i.d. sample of d points $\{N_i^{(k)}\}_{i\in[d]}$ from a categorical distribution with parameter $(p_1^{(k)}, \ldots, p_R^{(k)})$. Let

$$\hat{\lambda}_r^{(k)} = \frac{1}{d}\sum_i \mathbb{1}\{N_i^{(k)} \leq r\}, \quad \textit{for all } r \in [R]$$

denote the empirical cumulative proportions based on the d trials. Consider the estimator $\hat{f}$:

$$\hat{f}(\boldsymbol{x}) = \sum_{\boldsymbol{r}\in[R]^K} c_{\boldsymbol{r}} \mathbb{1}\left\{\boldsymbol{x} \in \underbrace{(\hat{\lambda}_{r_1-1}^{(1)},\ \hat{\lambda}_{r_1}^{(1)}] \times \cdots \times (\hat{\lambda}_{r_K-1}^{(K)},\ \hat{\lambda}_{r_K}^{(K)}]}_{=:\hat{I}_{\boldsymbol{r}}}\right\}, \quad \textit{for all } \boldsymbol{x} \in [0,1]^K.$$

Then

$$\mathbb{E}\left(\int_{\boldsymbol{x}\in[0,1]^K} |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})|^2 dx\right) \le \frac{8KR}{\sqrt{N}},$$

where the expectation is taken with respect to $\{N_i^{(k)}\}$.

Proof. Note that

$$\mathbb{E}\left(\int_{\boldsymbol{x}\in[0,1]^K} |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})|^2 dx\right) \le 4\mathbb{E}\int_{\boldsymbol{x}\in[0,1]^K} \mathbb{1}\{f(\boldsymbol{x}) \neq \hat{f}(\boldsymbol{x})\} d\boldsymbol{x}.$$

Therefore, it suffices to evaluate the Lebesgue measure of $\{\boldsymbol{x}\colon f(\boldsymbol{x}) \neq \hat{f}(\boldsymbol{x})\}$. Note that f and f' are the same over $I_{\boldsymbol{r}} \cap \hat{I}_{\boldsymbol{r}}$ for all $\boldsymbol{r}$. This implies

$$\begin{aligned}\{\boldsymbol{x}\colon f(\boldsymbol{x}) \neq f'(\boldsymbol{x})\} &\subset \cup_{\boldsymbol{r}}\{I_{\boldsymbol{r}} \Delta \hat{I}_{\boldsymbol{r}}\} \\ &\subset \cup_k \cup_r \left\{\boldsymbol{x}\colon x_k \in (\lambda_{r-1}^{(k)},\ \lambda_r^{(k)}] \Delta (\hat{\lambda}_{r-1}^{(k)},\ \hat{\lambda}_r^{(k)}]\right\},\end{aligned}$$

where the operation Δ denotes the symmetric difference between two sets, and the second line comes from the property for Cartesian product of intervals. Then,

$$\begin{aligned}\left|\{\boldsymbol{x}\colon f(\boldsymbol{x}) \neq f'(\boldsymbol{x})\}\right| &\le \sum_{k\in[K]} \sum_{r\in[R]} \left|(\lambda_{r-1}^{(k)},\ \lambda_r^{(k)}] \Delta (\hat{\lambda}_{r-1}^{(k)},\ \hat{\lambda}_r^{(k)}]\right| \\ &\le 2\sum_{k\in[K]} \sum_{r\in[R]} |\lambda_r^{(k)} - \hat{\lambda}_r^{(k)}|\end{aligned} \tag{7}$$

Note that by definition, $d\hat{\lambda}_r^{(k)} = \sum_i \mathbb{1}\left\{N_i^{(k)} \le r\right\}$ follows from Binomial distribution with parameters $(d, \lambda_r^{(k)})$. Therefore,

$$\mathbb{E}|\lambda_r^{(k)} - \hat{\lambda}_r^{(k)}|^2 = \frac{\lambda_r^{(k)}(1-\lambda_r^{(k)})}{d} \le \frac{1}{d}. \tag{8}$$

Plugging (8) into (7), we obtain

$$\begin{aligned}\left|\{\boldsymbol{x}\colon f(\boldsymbol{x}) \neq \hat{f}(\boldsymbol{x})\}\right|^2 &\le 4KR \sum_{k\in[K]} \sum_{r\in[R]} |\lambda_r^{(k)} - \hat{\lambda}_r^{(k)}|^2 \\ &\le \frac{4K^2R^2}{d}\end{aligned}$$

Henceforth,

$$\mathbb{E}\left|\{\boldsymbol{x}\colon f(\boldsymbol{x}) \neq \hat{f}(\boldsymbol{x})\}\right| \le \frac{2KR}{\sqrt{d}}.$$

□

5 Minimax lower bound

Theorem 5.1 (Minimax)**.** *For stochastic tensor model, there exists a constant $C > 0$ such that*

$$\inf_{\hat{\Theta}} \sup_{\Theta \in \mathcal{P}(R)} \mathbb{P}\left\{ \|\hat{\Theta} - \Theta\|_F > C\frac{\sigma^2}{p}\left(\frac{R^K + Kd\log R}{d^K}\right)\right\} > 0.2.$$

Proof. Nonparametric rate. We construct fixed $\boldsymbol{M}^{(k)} \in [0,1]^{d\times R}$ for all $k \in [K]$ as follows. For each $k \in [K]$, $\boldsymbol{M}\colon [d] \mapsto [R]$ partition the set $[d]$ into R equal-sized clusters, and we denote the mapping rule $\boldsymbol{M}(i) = \lceil \frac{iR}{d} \rceil$ for $i \in [d]$. For any binary tensor $\mathcal{C} \in \{0,1\}^{r_1\times\cdots\times r_K}$, we define the core tensor

$$\check{\mathcal{C}} = c\sqrt{\frac{\sigma^2 r^k}{pd^k}}\mathcal{C}.$$

We identify the tensor in $\{0,1\}^{r_1\times\cdots\times r_K}$ by vectors in $\{0,1\}^{r^K}$. By Lemma .., there exists some set Ω such that $|\Omega| \geq \exp\left(\frac{r^K}{4}\right)$ and $H(\mathcal{C},\mathcal{C}') \geq \frac{r^K}{4}$ for any $\mathcal{C} \neq \mathcal{C}' \in \Omega$. We consider the subspace in the original tensor induced by \$ $\square$

Definition 9 (Clustering function)**.** Let $M\colon [d] \mapsto [R]$ denotes a clustering function, where we use $M(i) \in [R]$ denote the cluster label to which the i was assigned. We use $M^{-1}(r) = \{d\colon M(d) = r\} \subset [d]$ to denote the set of indices that was assigned to cluster r. For notational convenience, we use $M \in [R]^d$ or $M \in \{0,1\}^{d\times R}$ exchangeable to denote the collection of all clustering functions that map $[d]$ to $[R]$.

Lemma 4. *There exists a subset $\Omega \in [R]^d$, such that $|\Omega| \geq \exp(d\log R/2)$ and*

$$H(\omega,\omega') \stackrel{def}{=} \left|\left\{d\colon \omega(d) \neq \omega'(d)\right\}\right| \geq \frac{d}{4}, \quad \textit{for all } \omega \neq \omega' \in \Omega.$$

Proof. Define

$$\Omega = \left\{\omega \in [R]^d\colon |\omega^{-1}(i)| = \frac{d}{R} \text{ for } i \in [R]\right\},$$

that is, Ω is the collection of clustering functions that generates the equal-sized clusters. Given any $\omega \in \Omega$, define its ε-neighborhood by

$$B(\omega,\varepsilon) = \left\{\omega' \in \Omega\colon H(\omega,\omega') \leq \varepsilon\right\}.$$

The packing number of Ω

$$\mathcal{M}(\varepsilon,\Omega,H) \geq \frac{|\Omega|}{\max_{\omega\in\Omega}|B(\omega,\varepsilon)|}.$$

Taking $\varepsilon = \frac{d}{4}$ gives

$$|B(\omega,\varepsilon)| \leq \binom{d}{d/4}R^{d/4} \leq (4e)^{d/4}R^{d/4} \leq \exp\left(\frac{1}{4}d\log R\right),$$

where we have used the inequality $\binom{d}{k} \leq (\frac{ed}{k})^k$ for all $k \leq d$. By stirling's formula

$$|\Omega| = \frac{d!}{[(d/R)!]^R} \approx \frac{\sqrt{d}\left(\frac{d}{e}\right)^d}{\sqrt{\frac{d}{R}}\left(\frac{d}{Re}\right)^d} \approx \sqrt{R}\,(R)^d \geq \exp\left(d\log R + o(d\log R)\right) \geq \exp\left(\frac{1}{2}d\log R\right).$$

Therefore,

$$\mathcal{M}\left(\frac{d}{4}, \Omega, H\right) \geq \exp\left(\frac{1}{2}d\log R\right).$$

$\square$

In Wang et al, the MSE for the tensor block model has the following asymptotical rate,

$$\mathrm{MSE}(\hat{\Theta}, \Theta) \stackrel{\text{def}}{=} \frac{1}{d^K} \|\Theta - \hat{\Theta}\|_F^2 \asymp \frac{R^K}{d^K} + \frac{K \log R}{d^{K-1}}, \quad \text{as } d \to \infty \text{ while fixing } R.$$

We investigate the asymptotic behavior as $R \asymp d^\delta$, for some $\delta \in [0, 1]$.

$$\mathrm{MSE}(\hat{\Theta}, \Theta) = \begin{cases} d^{-K}, & R = 1, \\ d^{-(K-1)}, & \delta = 0 \text{ (i.e. constant } R = \mathcal{O}(1)), \\ d^{-(K-1)} \log d, & \delta \in (0, \frac{1}{K}], \\ d^{-K(1-\delta)}, & \delta \in (\frac{1}{K}, 1]. \end{cases}$$

In particular, in the matrix case when $K = 2$, the asymptotic rate is $\mathcal{O}(d^{-1} \log d)$ whenever $R \asymp \mathcal{O}(\sqrt{d})$. This is faster than regular low-rank matrix decomposition but slower than Gao's paper.