Hypergraphon estimation error 4

Chanwoo Lee
March 7, 2021

1 Notation and problem setting

Let E be a set of possible m-uniform hyperedges from n vertices without diagonal entries,

$$E = \{(i_1, \dots, i_m) \in [n]^m \colon |\{i_1, \dots, i_m\}| = m\}.$$

We denote an index of m-uniform hyperedges as $\boldsymbol{\omega} = (\omega_1, \dots, \omega_m) \in [n]^m$ and a membership vector of m-vertices as $\boldsymbol{a} = (a_1, \dots, a_m) \in [k]^m$. Let $z \colon [n] \to [k]$ be a membership function. For a given membership function z and a membership vector $\boldsymbol{a} \in [k]^m$, define $E_{Z^{-1}(\boldsymbol{a})}$ as a set of m-uniform hyperedges whose clustering group belongs to $\boldsymbol{a}$ i.e.,

$$E_{z^{-1}(\boldsymbol{a})} = \{\boldsymbol{\omega} \in E \colon z(\omega_\ell) = a_\ell \text{ for all } \ell \in [m]\}.$$

We define a block average on a set $E_{z^{-1}(\boldsymbol{a})}$ for a given membership function z, a membership vector $\boldsymbol{a}$, and a tensor $\Theta \in ([n])^{\otimes m}$ as

$$\bar{\Theta}_{\boldsymbol{a}}(z) = \frac{1}{|E_{z^{-1}(\boldsymbol{a})}|} \sum_{\boldsymbol{\omega} \in E_{z^{-1}(\boldsymbol{a})}} \Theta_{\boldsymbol{\omega}}.$$

Now we consider an undirected m-uniform hypergraph. The connectivity is encoded by an adjacency tensor $\{\mathcal{A}_{\boldsymbol{\omega}}\}_{\boldsymbol{\omega} \in E}$ which takes values in $\{0, 1\}$. We assume that $\mathcal{A}_{\boldsymbol{\omega}} \sim$ Bernoulli$(\Theta_{\boldsymbol{\omega}})$, where

$$\Theta_{\boldsymbol{\omega}} = f(\xi_{\omega_1}, \dots, \xi_{\omega_m}), \text{ for all } \boldsymbol{\omega} = (\omega_1, \dots, \omega_m) \in E, \tag{1}$$

where $f \colon [0, 1]^m \to [0, 1]$ is a symmetric function called graphon such that $f(\xi_{\omega_1}, \dots, \xi_{\omega_m}) = f(\xi_{\sigma(\omega_1)}, \dots, \xi_{\sigma(\omega_m)})$ for all permutation $\sigma \colon [n] \to [n]$. Conventionally, we set $\Theta_{\boldsymbol{\omega}} = 0$ for all $\boldsymbol{\omega} \in [n]^m \setminus E$. $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ is a random vector of unobserved (latent) i.i.d. random variables uniformly distributed on $[0, 1]$. In this note, we further assume that a graphon f is α-Hölder continuous with a constant L.

Definition 1. For $\alpha \in (0, 1]$ and $L > 0$, a function $f \colon [0, 1]^m \to [0, 1]$ is a α-Hölder continuous with a constant L, denoted as $f \in \mathcal{H}(\alpha, L)$, if

$$|f(x) - f(y)| \le L \|x - y\|_\alpha^\alpha \quad \text{for all } x, y \in [0, 1]^m, \tag{2}$$

where the norm $\|x\|_p^p := \sum_{i=1}^m |x_i|^p$ for $x \in \mathbb{R}^m$.

2 Probability matrix estimation

2.1 Under hypergraphon model

Lemma 1 (Block approximation)**.** Suppose the true parameter Θ admits the form (1) with $f \in \mathcal{H}(\alpha, L)$. For every integer $k \le n$, there exists $z^* \colon [n] \to [k]$, satisfying

$$\frac{1}{|E|} \sum_{\boldsymbol{a} \in [k]^m} \sum_{\boldsymbol{\omega} \in E_{(z^*)^{-1}(\boldsymbol{a})}} (\Theta_{\boldsymbol{\omega}} - \bar{\Theta}_{\boldsymbol{a}}(z^*))^2 \le m^2 L^2 \left(\frac{1}{k^2}\right)^\alpha.$$

put all proofs in a seperate section.

Proof. Define $z^*\colon [n] \to [k]$ by

$$(z^*)^{-1}(\ell) = \left\{ i \in [n] \colon \xi_i \in \left[\frac{\ell - 1}{k}, \frac{\ell}{k} \right) \right\}, \quad \text{for each } \ell \in [k].$$

By the construction of z^* for $\xi_{\omega_\ell} \in [(a_{\ell-1} - 1)/k, a_\ell/k]$,

$$\begin{aligned} |f(\xi_{\omega_1}, \dots, \xi_{\omega_m}) - \bar{\Theta}_{\boldsymbol{a}}(z^*)| &= \left| f(\xi_{\omega_1}, \dots, \xi_{\omega_m}) - \frac{1}{|E_{z^{-1}(\boldsymbol{a})}|} \sum_{(\omega_1', \dots, \omega_m') \in E_{z^{-1}(\boldsymbol{a})}} f(\xi_{\omega_1'}, \dots, \xi_{\omega_m'}) \right| \\ &\le \frac{1}{|E_{z^{-1}(\boldsymbol{a})}|} \sum_{(\omega_1', \dots, \omega_m') \in E_{z^{-1}(\boldsymbol{a})}} |f(\xi_{\omega_1}, \dots, \xi_{\omega_m}) - f(\xi_{\omega_1'}, \dots, \xi_{\omega_m'})| \\ &\le \frac{1}{|E_{z^{-1}(\boldsymbol{a})}|} \sum_{(\omega_1, \dots, \omega_m) \in E_{z^{-1}(\boldsymbol{a})}} L \|(\xi_{\omega_1}, \dots, \xi_{\omega_m}) - (\xi_{\omega_1'}, \dots, \xi_{\omega_m'})\|_\alpha^\alpha \\ &\le mL \left(\frac{1}{k} \right)^\alpha. \end{aligned}$$

□

Let $\tilde{\Theta}$ be a minimizer of the least square error from the adjacency tensor $\mathcal{A}$,

$$\tilde{\Theta} = \operatorname*{arg\,min}_{\Theta \in \mathcal{P}_k} \sum_{\boldsymbol{\omega} \in E} (\mathcal{A}_{\boldsymbol{\omega}} - \Theta_{\boldsymbol{\omega}})^2,$$

where

$$\begin{aligned} \mathcal{P}_k = \{ \Theta \in ([0,1]^n)^{\otimes m} \colon \Theta = \mathcal{C} \times_2 \boldsymbol{M} \times_2 \cdots \times_m \boldsymbol{M}, \text{ with a} \\ \text{membership matrix } \boldsymbol{M} \text{ and a core tensor } \mathcal{C} \in ([0,1]^k)^{\otimes m} \}. \end{aligned}$$

We estimate the probability tensor by $\hat{\Theta} = \operatorname{cut}(\tilde{\Theta})$ such that

$$\operatorname{cut}(\Theta_{\boldsymbol{\omega}}) = \begin{cases} \Theta_{\boldsymbol{\omega}} & \text{if } \boldsymbol{\omega} \in E, \\ 0 & \text{if } \boldsymbol{\omega} \in [n]^m \setminus E. \end{cases} \tag{3}$$

Notice $\|\mathcal{A} - \hat{\Theta}\|_F^2 \le \|\mathcal{A} - \Theta\|_F^2$ for any k block tensor $\Theta \in \operatorname{cut}(\mathcal{P}_k)$.

Theorem 2.1 (Hypergraphon model)**.** Suppose the true parameter Θ admits the form (1) with $f \in \mathcal{H}(\alpha, L)$ Let $\hat{\Theta}$ be the estimator from (3). Then, there exist two constants $C_1, C_2 > 0$ such that,

$$\frac{1}{n^m} \|\hat{\Theta} - \Theta\|_F^2 \le C_1 \left(n^{\frac{-2m\alpha}{m+2\alpha}} + \frac{\log n}{n^{m-1}} \right),$$

with probability at least $1 - \exp\left(-C_2 \left(n \log n + n^{\frac{m^2}{m+2\alpha}} \right) \right)$ uniformly over $f \in \mathcal{H}(\alpha, L)$.

Remark 1. There are two ingredients in the rate of convergence, the nonparametric rate $n^{-2m\alpha/(m+2\alpha)}$ and the clustering rate $\log n / n^{m-1}$. Depending on constants m and α, convergence rate becomes

$$n^{\frac{-2m\alpha}{m+2\alpha}} + \frac{\log n}{n^{m-1}} \asymp \begin{cases} n^{\frac{-2\alpha}{1+\alpha}} & m = 2, \alpha \in (0,1), \\ \log n / n & m = 2, \alpha = 1, \\ n^{\frac{-2m\alpha}{m+2\alpha}} & m > 2. \end{cases}$$

Notice that the nonparametric rate dominates the clustering rate when $m > 2$. The intuitive explana-

the same —> roughly the same. Precisely, increase by a factor \frac{m}{m+2*alpha}

tion for this phenomenon is that the number of parameters increases exponentially due to the possible m-combinations from k clusters (which contributes to the nonparametric rate) while the possible number of k-cluster of n-vertices remains the same (which contributes to the clustering rate) when m increases given the choice of $k = \lceil n^{\frac{m}{m+2\alpha}} \rceil$. That means that the hypergraphon estimation becomes harder due to the increased number of parameters while cluster estimation becomes easier because we have relatively fixed combinations of clusters among n vertices ~~with more information about interaction among clusters~~. When $m = 2$, Theorem 2.1 reduces to the results in Gao et al. [2015]. with increased sample size

Remark 2. When we consider the vector case when $m = 1$, our model reduces to the one-dimensional regression problem such that

$$y_\omega \sim \text{Bernoulli}(\Theta_\omega), \quad \text{where } \Theta_\omega = f(\xi_\omega) \quad \text{for } \omega \in [n].$$

In this case, Therem 2.1 becomes

$$\frac{1}{n}\sum_{\omega=1}^{n}(\hat{\Theta}_\omega - \Theta_\omega)^2 \le C_1\left(n^{\frac{-2\alpha}{2\alpha+1}} + \log n\right). \tag{4}$$

The nonparametric rate $n^{-2\alpha/(2\alpha+1)}$ in (4) matches exactly with the nonparametric minimax estimation bound for Hölder class with smoothness α when we estimate f from the pairs $\{(\xi_i, y_i)\}_{\omega=1}^n$. However, the explanatory variables $\{\xi_i\}_{i=1}^n$ is not observed, we get an estimation error no better than a constant bound. The clustering rate $\log n$ in (4) which is slightly larger than a constant reflects this phenomenon.

Proof. First, we can find a block tensor Θ^* close to true Θ by Lemma 1. By triangular inequality,

$$\|\hat{\Theta} - \Theta\|_F^2 \le 2\underbrace{\|\hat{\Theta} - \Theta^*\|_F^2}_{\text{(i)}} + 2\underbrace{\|\Theta^* - \Theta\|_F^2}_{\text{(ii)}}. \tag{5}$$

Since we have already shown the error bound of (ii) in Lemma 1, we bound the error from (i). From the definition of $\hat{\Theta}$, we have

$$\|\hat{\Theta} - \mathcal{A}\|_F^2 \le \|\Theta^* - \mathcal{A}\|_F^2. \tag{6}$$

Combining (6) with the fact

$$\begin{aligned}\|\hat{\Theta} - \mathcal{A}\|_F^2 &= \|\hat{\Theta} - \Theta^* + \Theta^* - \mathcal{A}\|_F^2 \\ &= \|\hat{\Theta} - \Theta^*\|_F^2 + \|\Theta^* - \mathcal{A}\|_F + 2\langle\hat{\Theta} - \Theta^*, \Theta^* - \mathcal{A}\rangle,\end{aligned}$$

yields

$$\begin{aligned}\|\hat{\Theta} - \Theta^*\|_F^2 &\le 2\langle\hat{\Theta} - \Theta^*, \mathcal{A} - \Theta^*\rangle \\ &= 2\left(\langle\hat{\Theta} - \Theta^*, \mathcal{A} - \Theta\rangle + \langle\hat{\Theta} - \Theta^*, \Theta - \Theta^*\rangle\right) \\ &\le 2\|\hat{\Theta} - \Theta^*\|_F\left(\left\langle\frac{\hat{\Theta} - \Theta^*}{\|\hat{\Theta} - \Theta^*\|_F}, \mathcal{A} - \Theta\right\rangle + \|\Theta - \Theta^*\|_F\right).\end{aligned}$$

Let $\mathcal{M} = \{\boldsymbol{M} : \boldsymbol{M} \text{ is the collection of membership matrices}\}$. Then,

$$\begin{aligned}\left\langle\frac{\hat{\Theta} - \Theta^*}{\|\hat{\Theta} - \Theta^*\|_F}, \mathcal{A} - \Theta\right\rangle &\le \sup_{\Theta' \in \mathcal{P}_k}\sup_{\Theta'' \in \mathcal{P}_k}\left\langle\frac{\text{cut}(\Theta') - \text{cut}(\Theta'')}{\|\text{cut}(\Theta') - \text{cut}(\Theta'')\|_F}, \mathcal{A} - \Theta\right\rangle \\ &\le \sup_{\boldsymbol{M},\boldsymbol{M}' \in \mathcal{M}}\sup_{\mathcal{C},\mathcal{C}' \in ([0,1]^k)^{\otimes m}}\left\langle\frac{\text{cut}(\Theta(\boldsymbol{M},\mathcal{C})) - \text{cut}(\Theta(\boldsymbol{M}',\mathcal{C}'))}{\|\text{cut}(\Theta(\boldsymbol{M},\mathcal{C})) - \text{cut}(\Theta(\boldsymbol{M}',\mathcal{C}'))\|_F}, \mathcal{A} - \Theta\right\rangle.\end{aligned}$$

Notice that $\mathcal{A} - \Theta$ is sub-Gaussian with proxy parameter $\sigma^2 = 1/4$. By Lemma 7, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{M} \in \mathcal{M}} \sup_{\mathcal{C} \in ([-1,1]^k)^{\otimes m}} \left\langle \frac{\mathrm{cut}(\Theta(\boldsymbol{M}, \mathcal{C}))}{\|\mathrm{cut}(\Theta(\boldsymbol{M}, \mathcal{C}))\|_F}, \mathcal{A} - \Theta \right\rangle \geq t \right) \leq \exp(-C_1 t^2 + C_2 k^m + 2n \log k),$$

for some universal constants $C_1, C_2 > 0$. Choosing $t = C\sqrt{n \log k + k^m}$ yields

$$\text{(ii)} \leq C_1 \left(\left(\frac{1}{k} \right)^{2\alpha} + \left(\frac{k}{n} \right)^m + \frac{\log k}{n^{m-1}} \right),$$

with probability at least $1 - \exp\left(-C_2(n \log k + k^m)\right)$. Combinations of two error bounds in (5) and setting $k = \lceil n^{\frac{m}{m+2\alpha}} \rceil$, completes the theorem. □

2.2 Under stochastic block model

switch the order of 2.1 <—> 2.2.
Then no need to present the proof twice.

Theorem 2.2 (Stochastic block model)**.** Let $\hat{\Theta}$ be the estimator from (3). Suppose true probability tensor $\Theta \in \mathrm{cut}(\mathcal{P}_k)$ for fixed block size k. Then, there exists two constants $C_1, C_2 > 0$, such that

$$\frac{1}{n^m} \|\hat{\Theta} - \Theta\|_F^2 \leq C_1 \left(\frac{k}{n} \right)^m + \frac{\log k}{n^{m-1}},$$

with probability at least $1 - \exp\left(-C_2(n \log k + k^m)\right)$. In particular, suppose $k \asymp n^\delta$ for some $\delta \in [0, 1]$. Then, the convergence rate becomes

$$\left(\frac{k}{n} \right)^m + \frac{\log k}{n^{m-1}} \asymp \begin{cases} n^{-m} & k = 1, \\ n^{-m+1} & \delta = 0, k \geq 2, \\ n^{-m+1} \log(n) & \delta \in (0, 1/m], \\ n^{-m(1-\delta)} & \delta \in (1/m, 1]. \end{cases}$$

Remark 3. The first part of error is the nonparametric rate, which is determined by the number of parameters and the number of observation of the model. For the stochastic block model with k-clusters, the number of parameters is $\mathcal{O}(k^m)$ while the number of observation is $\mathcal{O}(n^m)$. The proportion of two numbers is absorbed in the nonparametric rate. The second part of error is the clustering rate. The number of possible k-clusters of n-vertices is $\mathcal{O}(k^n)$. We can check that the the possible number of clusters is reflected by logarithm in the clustering rate.

Proof. By similar way in the proof of Theorem 2.1, we have

$$\begin{aligned} \|\hat{\Theta} - \Theta\|_F^2 &\leq 2\langle \hat{\Theta} - \Theta, \mathcal{A} - \Theta \rangle \\ &= 2\|\hat{\Theta} - \Theta\|_F \left\langle \frac{\hat{\Theta} - \Theta}{\|\hat{\Theta} - \Theta\|_F}, \mathcal{A} - \Theta \right\rangle \\ &\leq 2\|\hat{\Theta} - \Theta\|_F \sup_{\Theta' \in \mathcal{P}_k} \sup_{\Theta'' \in \mathcal{P}_k} \left\langle \frac{\mathrm{cut}(\Theta') - \mathrm{cut}(\Theta'')}{\|\mathrm{cut}(\Theta') - \mathrm{cut}(\Theta'')\|_F}, \mathcal{A} - \Theta \right\rangle. \end{aligned}$$

Notice the last inequality holds because $\Theta \in \mathrm{cut}(\mathcal{P}_k)$. Therefore, we have the result following the proof of Theorem 2.1. □

3 Hypergraphon estimation

3.1 Hölder continuous graphon

For a given probability tensor Θ, define the empirical hypergraphon $f_\Theta \colon [0,1]^m \to [0,1]$ as the following piecewise constant function:

$$\hat{f}_\Theta(x_1, \dots, x_m) = \Theta_{\lfloor x_1 \rfloor, \dots, \lfloor x_m \rfloor}.$$

For any hypergraphon estimator $\hat{f}$, we define the squared error

$$\delta^2(\hat{f}, f) := \inf_{\tau \in \mathcal{T}} \int_{(0,1)^m} |f(\tau(\boldsymbol{x})) - \hat{f}(\boldsymbol{x})|^2 d\boldsymbol{x}, \tag{7}$$

where $\mathcal{T}$ is the set of all measure-preserving bijection $\tau \colon [0,1] \to [0,1]$. If there is no confusion, τ can be used as element-wise vector operation like (7).

Our goal is to construct the upper bound of error $\mathbb{E}\left[\delta^2(f_{\hat{\Theta}}, f)\right]$. By triangular inequality, we have

$$\mathbb{E}\left[\delta^2(f_{\hat{\Theta}}, f)\right] \le \underbrace{\frac{2}{n^m}\mathbb{E}\|\hat{\Theta} - \Theta\|_F^2}_{\text{(i)}} + 2\underbrace{\mathbb{E}\left[\delta^2(f_\Theta, f)\right]}_{\text{(ii)}}.$$

The following two lemmas show the upper bounds of (i) and (ii).

Lemma 2 (Estimation error for Hölder continuous hypergraphon)**.** Suppose the true parameter Θ admits the form (1) with $f \in \mathcal{H}(\alpha, L)$. Let $\hat{\Theta}$ be the estimator from (3). Then

Isn't Lemma 2 a direct corollary of Thm 2.1?

$$\frac{1}{n^m}\mathbb{E}\|\hat{\Theta} - \Theta\|_F^2 \le C\left(n^{\frac{-2m\alpha}{m+2\alpha}} + \frac{\log n}{n^{m-1}}\right),$$

From tail bound —> expectation bound

for some constant C only depending proportionally on L^2.

Let X be bounded r.v.

EX <= t P(X<=t) + L P(X>t) = t+(L-t)P(X>t). The first term dominates. So EX~t

Proof. Expectation is with respect to Bernoulli distribution $\mathcal{A}$ and uniform distributions $\boldsymbol{\xi} := \{(\xi_{\omega_1}, \dots, \xi_{\omega_m})\}_{\boldsymbol{\omega} \in E}$. Notice that

$$\frac{1}{n^m}\mathbb{E}_{\mathcal{A},\boldsymbol{\xi}}\|\hat{\Theta} - \Theta\|_F^2 = \frac{1}{n^m}\mathbb{E}_{\boldsymbol{\xi}}\left[\mathbb{E}_{\mathcal{A}}\left[\|\hat{\Theta} - \Theta\|_F^2 \middle| \boldsymbol{\xi}\right]\right]$$

Similar to the proof of Theorem 2.1, let a block tensor Θ^* be a block tensor satisfying Lemma 1. Given $\boldsymbol{\xi}$, we have the following by triangular inequality,

$$\|\hat{\Theta} - \Theta\|_F^2 \le 2\underbrace{\|\hat{\Theta} - \Theta^*\|_F^2}_{\text{(i)}} + 2\underbrace{\|\Theta^* - \Theta\|_F^2}_{\text{(ii)}}. \tag{8}$$

Since we have already shown the error bound of (ii) in Lemma 1 given $\boldsymbol{\xi}$, we bound the error from (i) given $\boldsymbol{\xi}$. From the definition of $\hat{\Theta}$, we have

We use this proof idea but write >=3 times. Possible to reduce to one?

$$\|\hat{\Theta} - \Theta^*\|_F^2 \le \left(2\left\langle \frac{\hat{\Theta} - \Theta^*}{\|\hat{\Theta} - \Theta^*\|_F}, \mathcal{A} - \Theta \right\rangle + 2\|\Theta - \Theta^*\|_F\right)^2 \le 4\left(\left\langle \frac{\hat{\Theta} - \Theta^*}{\|\hat{\Theta} - \Theta^*\|_F}, \mathcal{A} - \Theta \right\rangle^2 + \|\Theta - \Theta^*\|_F^2\right). \tag{9}$$

Notice $\|\Theta^* - \Theta\|_F$ is a constant with respect to $\mathcal{A}$. Therefore, only random variable related to $\mathcal{A}$ in $\|\hat{\Theta} - \Theta^*\|_F$

is the inner product term. Let $\mathcal{M} = \{\boldsymbol{M} : \boldsymbol{M} \text{ is the collection of membership matrices}\}$. Then,

$$\begin{aligned}
\left\langle \frac{\hat{\Theta} - \Theta^*}{\|\hat{\Theta} - \Theta^*\|_F}, \mathcal{A} - \Theta \right\rangle^2 &\leq \sup_{\Theta' \in \mathcal{P}_k} \sup_{\Theta'' \in \mathcal{P}_k} \left\langle \frac{\mathrm{cut}(\Theta') - \mathrm{cut}(\Theta'')}{\|\mathrm{cut}(\Theta') - \mathrm{cut}(\Theta'')\|_F}, \mathcal{A} - \Theta \right\rangle^2 \\
&\leq \sup_{\boldsymbol{M}, \boldsymbol{M}' \in \mathcal{M}} \sup_{\mathcal{C}, \mathcal{C}' \in ([0,1]^k)^{\otimes m}} \left\langle \frac{\mathrm{cut}(\Theta(\boldsymbol{M}, \mathcal{C})) - \mathrm{cut}(\Theta(\boldsymbol{M}', \mathcal{C}'))}{\|\mathrm{cut}(\Theta(\boldsymbol{M}, \mathcal{C})) - \mathrm{cut}(\Theta(\boldsymbol{M}', \mathcal{C}'))\|_F}, \mathcal{A} - \Theta \right\rangle^2 \\
&\leq \sup_{\boldsymbol{M}, \boldsymbol{M}' \in \mathcal{M}} \sup_{\mathcal{C}, \mathcal{C}' \in ([0,1]^k)^{\otimes m}} \left\langle \frac{\mathrm{cut}(\Theta(\boldsymbol{M}, \mathcal{C} - \mathcal{C}')) + \mathrm{cut}(\Theta(\boldsymbol{M} - \boldsymbol{M}', \mathcal{C}'))}{\|\mathrm{cut}(\Theta(\boldsymbol{M}, \mathcal{C} - \mathcal{C}')) + \mathrm{cut}(\Theta(\boldsymbol{M} - \boldsymbol{M}', \mathcal{C}'))\|_F}, \mathcal{A} - \Theta \right\rangle^2
\end{aligned}$$

Notice that $\mathcal{A} - \Theta$ is sub-Gaussian with proxy parameter $\sigma^2 = 1/4$. By Lemma 8, we have

$$\mathbb{E}_{\mathcal{A}} \left[\left\langle \frac{\hat{\Theta} - \Theta^*}{\|\hat{\Theta} - \Theta^*\|_F}, \mathcal{A} - \Theta \right\rangle^2 \middle| \boldsymbol{\xi} \right] \leq C \left(n \log k + k^m \right), \tag{10}$$

for some constant $C > 0$.

Combination of (8),(9),(10), and Lemma 1 yields

$$\frac{1}{n^m} \mathbb{E}_{\mathcal{A}} \left[\|\hat{\Theta} - \Theta\|_F^2 \middle| \boldsymbol{\xi} \right] \leq C \left(\left(\frac{1}{k} \right)^{2\alpha} + \left(\frac{k}{n} \right)^m + \frac{\log k}{n^{m-1}} \right), \tag{11}$$

for some constant $C > 0$. Setting $k = \lceil n^{\frac{m}{m+2\alpha}} \rceil$ and the fact that the inequality (11) is not depending on the random variable $\boldsymbol{\xi}$, completes the proof as follows.

$$\frac{1}{n^m} \mathbb{E}_{\mathcal{A}, \boldsymbol{\xi}} \|\hat{\Theta} - \Theta\|_F^2 = \frac{1}{n^m} \mathbb{E}_{\boldsymbol{\xi}} \left[\mathbb{E}_{\mathcal{A}} \left[\|\hat{\Theta} - \Theta\|_F^2 \middle| \boldsymbol{\xi} \right] \right] \leq C \left(n^{\frac{-2m\alpha}{m+2\alpha}} + \frac{\log n}{n^{m-1}} \right).$$

□

Lemma 3 (Agnostic error for Hölder continuous hypergraphon)**.** Suppose $f \in \mathcal{H}(\alpha, L)$. Then

$$\mathbb{E} \left[\delta^2(f_\Theta, f) \right] \leq \frac{C}{n^\alpha}, \qquad \text{write m^2L^2 in the numerator.}$$

for some constant $C > 0$ only depending proportionally on $m^2 L^2$.

Proof. By triangular inequality, we have

$$\mathbb{E} \left[\delta^2(f_\Theta, f) \right] \leq 2\mathbb{E} \left[\delta^2(f_\Theta, f_{\Theta'}) \right] + 2\mathbb{E} \left[\delta^2(f_{\Theta'}, f) \right],$$

where $\Theta' \in ([0,1]^n)^{\otimes m}$ such that $\Theta'_{\boldsymbol{\omega}} = f(\xi_{\omega_1}, \ldots, \xi_{\omega_m})$ for all $\boldsymbol{\omega} \in [n]^m$. Notice $\Theta'_{\boldsymbol{\omega}} = \Theta_{\boldsymbol{\omega}}$ for $\boldsymbol{\omega} \in E$ but $\Theta_{\boldsymbol{\omega}} = 0$ for $\boldsymbol{\omega} \in [n]^m \setminus E$. By definition of Θ',

$$\mathbb{E} \left[\delta^2(f_\Theta, f_{\Theta'}) \right] = \int_{[0,1]^m} |f_\Theta(\boldsymbol{x}) - f_{\Theta'}(\boldsymbol{x})| d\boldsymbol{x} < \frac{C}{n},$$

for some $C > 0$ only depending on m. This is because $|f_\Theta(\boldsymbol{x}) - f_{\Theta'}(\boldsymbol{x})| = 0$ outside of a set of measure $\left(n^m - \binom{n}{m} n!\right) / n^m = C/n$. Notice $C = \mathcal{O}(m^2)$. Hence it suffices to prove that

$$\mathbb{E} \left[\delta^2(f_{\Theta'}, f) \right] \leq \frac{C}{n^\alpha}.$$

We have

$$\delta^2(f_{\Theta'}, f) = \inf_{\tau\in\mathcal{T}} \sum_{i_1,\ldots,i_m=1}^{n} \int_{(i_1-1)/n}^{i_1/n} \cdots \int_{(i_m-1)/n}^{i_m/n} |f(\tau(x_1),\ldots,\tau(x_m)) - \Theta'_{i_1,\ldots,i_m}|^2 dx_1\cdots dx_m$$

The infimum over all measure-preserving bijection is smaller than the minimum over the subclass of measure-preserving bijection τ such that

$$\int_{(i_1-1)/n}^{i_1/n} \cdots \int_{(i_m-1)/n}^{i_m/n} f(\tau(x_1),\ldots\tau(x_m))dx_1\cdots dx_m = \int_{(\sigma(i_1)-1)/n}^{\sigma(i_1)/n} \cdots \int_{(\sigma(i_m)-1)/n}^{\sigma(i_m)/n} f(x_1,\ldots x_m)dx_1\cdots dx_m$$

for some permutation σ. For $x \in \prod_{\ell=1}^m [(\sigma(i_\ell)-1)/n, \sigma(i_\ell)/n]$,

$$\begin{aligned} |f(x_1,\ldots,x_m) - f(\xi_1,\ldots,\xi_m)|^2 \le & 2\left|f(x_1,\ldots,x_m) - f\left(\frac{\sigma(i_1)}{n+1},\ldots\frac{\sigma(i_m)}{n+1}\right)\right|^2 \\ & + 2\left|f\left(\frac{\sigma(i_1)}{n+1},\ldots\frac{\sigma(i_m)}{n+1}\right) - f(\xi_{(\sigma(i_1))},\ldots,\xi_{(\sigma(i_m))})\right|^2 \\ & + 2\left|f(\xi_{(\sigma(i_1))},\ldots,\xi_{(\sigma(i_m))}) - f(\xi_{i_1},\ldots,\xi_{i_m})\right|^2, \end{aligned} \tag{12}$$

where $\xi_{(\ell)}$ denotes the ℓ-th largest element of the set $\{\xi_1,\ldots,\xi_n\}$. Choose random permutation σ such that $\xi_{\sigma^{-1}(1)} \le \xi_{\sigma^{-1}(2)} \cdots \le \xi_{\sigma^{-1}(n)}$. Then the third summand in (12) is 0 almost surely.

For the first summand in (12), notice $(\sigma(i_1)/(n+1),\ldots\sigma(i_m)/(n+1)) \in \prod_{\ell=1}^m [(\sigma(i_\ell)-1)/n, \sigma(i_\ell)/n]$. From (2), we obtain

$$\left|f(x_1,\ldots,x_m) - f\left(\frac{\sigma(i_1)}{n+1},\ldots\frac{\sigma(i_m)}{n+1}\right)\right|^2 \le m^2M^2\left(\frac{1}{n}\right)^{2\alpha}.$$

Integrating and taking expectation on the first summand yields,

$$\mathbb{E}\left[\sum_{i_1,\ldots,i_m=1}^{n} \int_{(\sigma(i_1)-1)/n}^{\sigma(i_1)/n} \cdots \int_{(\sigma(i_m)-1)/n}^{\sigma(i_m)/n} \left|f(x_1,\ldots,x_m) - f\left(\frac{\sigma(i_1)}{n+1},\ldots\frac{\sigma(i_m)}{n+1}\right)\right|^2 dx_1\cdots dx_m\right] \le m^2L^2\left(\frac{1}{n}\right)^{2\alpha}. \tag{13}$$

With (2), the second summand on (12) is bounded,

$$\begin{aligned} \left|f\left(\frac{\sigma(i_1)}{n+1},\ldots\frac{\sigma(i_m)}{n+1}\right) - f(\xi_{(\sigma(i_1))},\ldots,\xi_{(\sigma(i_m))})\right|^2 &\le \left(L\sum_{\ell=1}^{m}\left|\frac{\sigma(i_\ell)}{n+1} - \xi_{(\sigma(i_\ell))}\right|^\alpha\right)^2 \\ &\le 2L^2\sum_{\ell=1}^{m}\left|\frac{\sigma(i_\ell)}{n+1} - \xi_{(\sigma(i_\ell))}\right|^{2\alpha}. \end{aligned}$$

Integrating and taking expectation on the second summand yields,

$$\begin{aligned} &\mathbb{E}\left[\sum_{i_1,\ldots,i_m=1}^{n} \int_{(\sigma(i_1)-1)/n}^{\sigma(i_1)/n} \cdots \int_{(\sigma(i_m)-1)/n}^{\sigma(i_m)/n} \left|f\left(\frac{\sigma(i_1)}{n+1},\ldots\frac{\sigma(i_m)}{n+1}\right) - f(\xi_{(\sigma(i_1))},\ldots,\xi_{(\sigma(i_m))})\right|^2 dx_1\cdots dx_m\right] \\ &\le \left[\sum_{i_1,\ldots,i_m=1}^{n} \int_{(\sigma(i_1)-1)/n}^{\sigma(i_1)/n} \cdots \int_{(\sigma(i_m)-1)/n}^{\sigma(i_m)/n} 2L^2\sum_{\ell=1}^{m}\left|\frac{\sigma(i_\ell)}{n+1} - \xi_{(\sigma(i_\ell))}\right|^{2\alpha} dx_1\cdots dx_m\right] \end{aligned}$$

$$
\begin{aligned}
&= 2mL^2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{i}{n+1}-\xi_{(i)}\right|^{2\alpha}\right] \\
&\le 2mL^2 \max_{i=1,\dots,n}\mathbb{E}\left[\left|\frac{i}{n+1}-\xi_{(i)}\right|^{2\alpha}\right] \\
&\le 2mL^2 \max_{i=1\dots,n}\left[\mathrm{Var}(\xi_{(i)})\right]^{\alpha} \le CmL^2\left(\frac{1}{n}\right)^{\alpha}, \qquad (14)
\end{aligned}
$$

where we have used $\mathbb{E}(\xi_{(\ell)}) = \ell/(n+1)$, $\mathrm{Var}(\xi_{(\ell)}) \le C/n$ and Jensen's inequality. Combining (13) and (14) proves the lemma. □

By Lemma 2 and Lemma 3, we have the following theorem.

Theorem 3.1 (Mean square error of smooth hypergraphon)**.** Suppose the true parameter Θ admits the form (1) with $f \in \mathcal{H}(\alpha, L)$. Let $\hat{\Theta}$ be the estimator from (3). Then, there exists a positive constant $C > 0$ only depending on m and L such that

$$
\mathbb{E}\left[\delta^2(f_{\hat{\Theta}}, f)\right] \le C\left(n^{\frac{-2m\alpha}{m+2\alpha}} + \frac{\log n}{n^{m-1}} + \frac{1}{n^{\alpha}}\right).
$$

Remark 4. The error rate is dominated by agnostic error $n^{-\alpha}$ for all combinations of $\{(m,\alpha)\}_{m\ge 2,\alpha\in(0,1]}$ except $(m,\alpha) = (2,1)$, when $\log n/n$ is the dominating error term. This agnostic error is unavoidable when considering mean square error because we estimate whole function values on $[0,1]^m$ from $\binom{n}{m} \asymp n^m$ sample size. Notice that the estimation error term getting smaller when m increases because the increased number of sample size is larger than increased number of parameters. However, the agnostic error almost remains the same regardless of m because the domain we have to estimate is increased to $[0,1]^m$, which has the same rate of increased sample size n^m. That explains why the agnostic error term dominates as m increases.

3.2 Piecewise constant hypergraphon

Define $\mathcal{F}(k)$ the collection of k-piecewise constant hypergraphons such that for some $\mathcal{C} \in ([0,1]^k)^{\otimes m}$ and some $\phi\colon [0,1] \to [k]$,

$$
f(x_1,\dots,x_m) = \mathcal{C}_{\phi(x_1),\dots,\phi(x_m)} \quad \text{for all } x_i \in [0,1], i = 1,\dots m.
$$

The estimation error associated to k-piecewise hypergraphon is bounded as follows.

latex format?

Lemma 4. [Estimation error for k-piecewise constant hypergraphon] Suppose the true parameter Θ admits the form (1) with $f \in \mathcal{F}(k)$. Let $\hat{\Theta}$ be the estimator from (3). Then

again, this is a direct corollary of Thm 2.2

$$
\frac{1}{n^m}\mathbb{E}\|\hat{\Theta}-\Theta\|_F^2 \le C\left(\left(\frac{k}{n}\right)^m + \frac{\log k}{n^{m-1}}\right),
$$

for some constant C only depending proportionally on L^2.

Proof. Expectation is with respect to Bernoulli distribution $\mathcal{A}$ and uniform distributions $\boldsymbol{\xi} := \{(\xi_{\omega_1},\dots,\xi_{\omega_m})\}_{\boldsymbol{\omega}\in E}$. Notice that

$$
\frac{1}{n^m}\mathbb{E}_{\mathcal{A},\boldsymbol{\xi}}\|\hat{\Theta}-\Theta\|_F^2 = \frac{1}{n^m}\mathbb{E}_{\boldsymbol{\xi}}\left[\mathbb{E}_{\mathcal{A}}\left[\|\hat{\Theta}-\Theta\|_F^2\middle|\boldsymbol{\xi}\right]\right]
$$

Similar to the proof of Therem 2.2, we have

$$
\mathbb{E}_{\mathcal{A}}\left[\|\hat{\Theta}-\Theta\|_F^2\middle|\boldsymbol{\xi}\right] \le 4\mathbb{E}_{\mathcal{A}}\left[\left\langle\frac{\hat{\Theta}-\Theta}{\|\hat{\Theta}-\Theta\|_F}, \mathcal{A}-\Theta\right\rangle^2\middle|\boldsymbol{\xi}\right]
$$

$$\leq 4\mathbb{E}_{\mathcal{A}}\left[\sup_{\Theta'\in\mathcal{P}_k}\sup_{\Theta''\in\mathcal{P}_k}\left\langle\frac{\mathrm{cut}(\Theta')-\mathrm{cut}(\Theta'')}{\|\mathrm{cut}(\Theta')-\mathrm{cut}(\Theta'')\|_F},\mathcal{A}-\Theta\right\rangle^2\middle|\boldsymbol{\xi}\right].$$

Notice the last inequality holds because $\Theta\in\mathrm{cut}(\mathcal{P}_k)$. Therefore, we have the result following the proof of Lemma 2. □

The egonostic error associated to k-piecewise hypergraphon is bounded as follows.

Lemma 5 (Agnostic error for k-piecewise constant hypergraphon)**.** Consider the hypergraphon model (1). For all integers $k\leq n$, $f\in\mathcal{F}(k)$, we have

$$\mathbb{E}[\delta^2(f_\Theta,f)]\leq C\sqrt{\frac{k}{n}},$$ write m^2 in the bound.

for some constant $C>0$ only depending proportionally on m^2.

Proof. By triangular inequality, we have

$$\mathbb{E}\left[\delta^2(f_\Theta,f)\right]\leq 2\mathbb{E}\left[\delta^2(f_\Theta,f_{\Theta'})\right]+2\mathbb{E}\left[\delta^2(f_{\Theta'},f)\right],$$

where $\Theta'\in([0,1]^n)^{\otimes m}$ such that $\Theta'_{\boldsymbol{\omega}}=f(\xi_{\omega_1},\ldots,\xi_{\omega_m})$ for all $\boldsymbol{\omega}\in[n]^m$. Notice $\Theta'_{\boldsymbol{\omega}}=\Theta_{\boldsymbol{\omega}}$ for $\boldsymbol{\omega}\in E$ but $\Theta_{\boldsymbol{\omega}}=0$ for $\boldsymbol{\omega}\in[n]^m\setminus E$. By definition of Θ',

$$\mathbb{E}\left[\delta^2(f_\Theta,f_{\Theta'})\right]=\int_{[0,1]^m}|f_\Theta(\boldsymbol{x})-f_{\Theta'}(\boldsymbol{x})|d\boldsymbol{x}<\frac{C}{n},$$

for some $C>0$ only depending on m. This is because $|f_\Theta(x)-f_{\Theta'}(x)|=0$ outside of a set of measure $\left(n^m-\binom{n}{m}n!\right)/n^m=C/n$. Notice $C=\mathcal{O}(m^2)$. Hence it suffices to prove that

$$\mathbb{E}\left[\delta^2(f_{\Theta'},f)\right]\leq C\sqrt{\frac{k}{n}}.$$

Without loss of generality, we assume that all the rows of the tensor $\{\mathcal{C}_{i,\cdots,\cdot}\}_{i=1}^k$ are distinct and $\mu_a:=\mu(\phi^{-1}(a))$ is positive for all $a\in[k]$, where μ is the Lebesgue measure. Otherwise, we can find $k'\leq k$ satisfying the above assumptions, so the problem is reduced to $f\in\mathcal{F}_{k'}$.

First, we construct an ordered hypergraphon f' such that $f'(x)=f(\tau(x))$ for all $x\in[0,1]^m$ and some measure-preserving bijection $\tau\colon[0,1]\to[0,1]$. For any $b\in[k]$, define the cumulative function

$$F_\phi(b)=\sum_{a=1}^{b}\mu_a,$$

and set $F_\phi(0)=0$. For any $\boldsymbol{a}=(a_1,\ldots,a_m)\in[k]^m$, define a grid $\Pi_{\boldsymbol{a}}(\phi)=\prod_{\ell=1}^m[F_\phi(a_\ell-1),F_\phi(a_\ell))$. Finally, we construct the ordered hypergraphon f' as

$$f'(\boldsymbol{x})=\sum_{a_1,\ldots,a_m=1}^{k}\mathcal{C}_{a_1,\ldots,a_m}\mathbb{1}_{\Pi_{\boldsymbol{a}}(\phi)}(\boldsymbol{x})\quad\text{for }\boldsymbol{x}\in[0,1]^m.$$

Second, we construct an ordered hypergraphon $f'_{\Theta'}$ such that $f'_{\Theta'}(x)=f_{\Theta'}(\tau(x))$ for all $x\in[0,1]^m$ and some measure-preserving bijection $\tau\colon[0,1]\to[0,1]$. Define the empirical frequency of group $a\in[k]$ as

$$\hat{\mu}_a=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\xi_i\in\phi^{-1}(a)\}},$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ is a random variable in the model (1). Consider a function $\psi\colon [0,1] \to [k]$ such that

(i) $\psi(x) = a$ for all $a \in [k]$ and $x \in \Big[F_\phi(a-1), F_\phi(a-1) + \min(\hat{\lambda}_a, \lambda_a)\Big)$

(ii) $\mu(\psi^{-1}(a)) = \hat{\lambda}_a$ for all $a \in [k]$.

Based on the function ψ, we constructing the ordered hypergraphon $f'_{\Theta'}$ in the view of (i) as

$$f'_{\Theta'}(\boldsymbol{x}) = \mathcal{C}_{\psi(x_1),\ldots,\psi(x_m)} \quad \text{for } \boldsymbol{x} = (x_1, \ldots, x_m) \in [0,1]^m.$$

By condition (ii), we guarantee there exists measure-preserving bijection τ such that $f_{\Theta'}(\tau(\boldsymbol{x})) = f'_{\Theta'}(\boldsymbol{x})$ for all $\boldsymbol{x} \in [0,1]^m$.

Now we show the existence of such function ψ satisfying both (i) and (ii). Consider the case when group $a \in [k]$ satisfies $\lambda_a > \hat{\lambda}_a$. Then, conditions (i) and (ii) holds trivially if we set $\phi^{-1}(a) = [F_\phi(a-1), F_\phi(a-1)+\hat{\lambda}_a)$. Notice that for this $\psi^{-1}(a)$, ψ does not assign points in the interval $[F_\phi(a-1)+\hat{\lambda}_a, F_\phi(a-1)+\lambda_a)$ the label a. Therefore, when $\lambda_a > \hat{\lambda}_a$, the total measure of unassigned points is $\sum_{a\colon \lambda_a > \hat{\lambda}_a} (\lambda_a - \hat{\lambda}_a)$. Now consider the case when $\lambda_a < \hat{\lambda}_a$. By condition (i), we must have $\psi(x) = a$ for $x \in [F_\phi(a-1), F_\phi(a-1)+\lambda_a)$, To satisfy condition (ii), we need additional points to be assigned as $\psi(x) = a$ and the measure of those points for each a is $\hat{\lambda}_a - \lambda_a$. Therefore, the total measure of additional points needed when $\lambda_a < \hat{\lambda}_a$ is $\sum_{a\colon \lambda_a < \hat{\lambda}_a} (\hat{\lambda}_a - \lambda_a)$. Since $\sum_{a=1}^k \lambda_a = \sum_{a=1}^k \hat{\lambda}_a = 1$, we have

$$\sum_{a\colon \lambda_a > \hat{\lambda}_a} (\lambda_a - \hat{\lambda}_a) = \sum_{a\colon \lambda_a < \hat{\lambda}_a} (\hat{\lambda}_a - \lambda_a),$$

which implies, we can successfully use a set of unlabeled points in the case $\lambda_a > \hat{\lambda}_a$ to satisfy condition (ii) in the case $\lambda_a < \hat{\lambda}_a$. Therefore, we can always find a function ψ satisfying conditions (i) and (ii).

We have constructed the ordered hypergraphons f' and $f'_{\Theta'}$ from f and $f_{\Theta'}$ such that

$$\delta^2(f_{\Theta'}, f) = \delta^2(f'_{\Theta'}, f').$$

Finally, we have

$$\begin{aligned}
\delta^2(f'_{\Theta'}, f') &\le \int_{[0,1]^m} |f'_{\Theta'}(\boldsymbol{x}) - f'(\boldsymbol{x})|^2 d\boldsymbol{x} \\
&\le \int_{[0,1]^m} \mathbb{1}_{\{f'_{\Theta'}(\boldsymbol{x}) \neq f'(\boldsymbol{x})\}} d\boldsymbol{x}
\end{aligned}$$

The two functions $f'_{\Theta'}(\boldsymbol{x})$ and $f'(\boldsymbol{x})$ are equal except the case when

$$\boldsymbol{x} \in \Big\{\boldsymbol{x} \in [0,1]^m \colon \exists \ell \in [m] \text{ such that } x_\ell \in \Big[F_\phi(a-1) + \min(\hat{\lambda}_a, \lambda_a), F_\phi(a-1) + \lambda_a\Big) \text{ for some } a \in [k]\Big\}.$$

Thus, we have

$$\delta^2(f_{\Theta'}, f) \le m \sum_{a=1}^k |\lambda_a - \hat{\lambda}_a|.$$

Finally, we have

$$\begin{aligned}
\mathbb{E}\left[\delta^2(f_{\Theta'}, f)\right] &\le m \sum_{a=1}^k \mathbb{E}|\lambda_a - \hat{\lambda}_a| \\
&\le \sum_{a=1}^k m \sqrt{\mathbb{E}|\lambda_a - \hat{\lambda}_a|^2} \qquad \text{[Jensen's inequality]}
\end{aligned}$$

$$\leq m \sum_{a=1}^{k} \sqrt{\frac{\lambda_a}{n}} \qquad [n\hat{\lambda}_a \sim \mathrm{Binom}(n, \lambda_a)]$$
$$\leq m\sqrt{\frac{k}{n}}$$

where the last inequality is from Cauchy-Schwarz inequality. □

By Lemma 4, 5, and triangular inequality, we have the following theorem.

Theorem 3.2 (Mean square error of k-piecewise constant hypergraphon)**.** Suppose the true parameter Θ admits the form (1) with $f \in \mathcal{F}_k$. Let $\hat{\Theta}$ be the estimator from (3). Then, there exists a positive constant $C > 0$ only depending on m and L such that

$$\mathbb{E}\left[\delta^2(f_{\hat{\Theta}}, f)\right] \leq C\left(\left(\frac{k}{n}\right)^m + \frac{\log k}{n^{m-1}} + \sqrt{\frac{k}{n}}\right).$$

Remark 5. The error rate is dominated by agnostic error $\sqrt{k/n}$ for all combinations of $\{(m, \alpha)\}_{m\geq 2, \alpha\in(0,1]}$. We have the similar argument in Remark 4.

4. Proofs. put all proofs here.

4 Technical lemmas

Lemma 6 (Gaussian maxima under full embedding)**.** Let $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$ be a deterministic matrix with rank $r \leq \min(d_1, d_2)$. Let $\boldsymbol{y} \in \mathbb{R}^{d_2}$ 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

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

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 $\boldsymbol{c} \in \mathbb{R}^r$, we have

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

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

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

where we define $\boldsymbol{x} = (\boldsymbol{u}_1^T\boldsymbol{y}, \ldots, \boldsymbol{u}_r^T\boldsymbol{y})^T \in \mathbb{R}^r$. 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. □

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

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

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

Lemma 7 (Sub-Gaussian tail bound)**.**

$$\mathbb{P}\left(\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\mathcal{C}_1,\mathcal{C}_2\in([0,1]^k)^{\otimes m}}\left\langle\frac{\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))}{\|\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))\|_F},\mathcal{A}-\Theta\right\rangle\geq t\right)\leq\exp(-C_1t^2+C_2k^m+2n\log k),$$

for some constant $C_1, C_2 > 0$.

Proof. Define $\Omega \subset [n^m]$, a collection of indices of the vectorized tensor corresponding to diagonal entires of the original tensor, i.e., for all $\boldsymbol{\omega} \in [n]^m \setminus E$, there exists $\omega' \in \Omega$ such that $\Theta_{\boldsymbol{\omega}} = \mathrm{vec}(\Theta)_{\omega'}$. Notice that

$$\mathrm{vec}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1)-\Theta(\boldsymbol{M}_2,\mathcal{C}_2))=\underbrace{\begin{bmatrix}\boldsymbol{M}_1^{\otimes m} & -\boldsymbol{M}_2^{\otimes m}\end{bmatrix}}_{=:\boldsymbol{A}\in\{0,1\}^{n^m\times 2k^m}}\underbrace{\begin{bmatrix}\mathrm{vec}(\mathcal{C}_1)\\ \mathrm{vec}(\mathcal{C}_2)\end{bmatrix}}_{=:\boldsymbol{c}\in[0,1]^{2k^m\times 1}}.$$

Defining a mapping $\mathrm{mcut}\colon \mathbb{R}^{n^m\times 2k^m}\to\mathbb{R}^{n^m\times 2k^m}$, such that

$$\mathrm{mcut}(\boldsymbol{A})_{i\cdot}=\begin{cases}\boldsymbol{A}_{i\cdot} & \text{if } i\in[n^m]\setminus\Omega,\\ 0 & \text{if } i\in\Omega.\end{cases}$$

gives us

$$\mathrm{vec}\left(\mathrm{cut}(\Theta(M_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(M_2,\mathcal{C}_2))\right)=\tilde{\boldsymbol{A}}\boldsymbol{c},$$

where $\tilde{\boldsymbol{A}} := \mathrm{mcut}(\boldsymbol{A})$. Based on the above notations, we have

$$\begin{aligned}&\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\mathcal{C}_1,\mathcal{C}_2\in([0,1]^k)^{\otimes m}}\left\langle\frac{\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))}{\|\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))\|_F},\mathcal{A}-\Theta\right\rangle\\ &\qquad=\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\mathcal{C}_1,\mathcal{C}_2\in([0,1]^k)^{\otimes m}}\left\langle\frac{\mathrm{vec}\left(\mathrm{cut}(\Theta(M_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(M_2,\mathcal{C}_2))\right)}{\|\mathrm{vec}\left(\mathrm{cut}(\Theta(M_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(M_2,\mathcal{C}_2))\right)\|_F},\mathrm{vec}\left(\mathcal{A}-\Theta\right)\right\rangle\\ &\qquad\leq\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\boldsymbol{c}\in\mathbb{R}^{2k^m}}\left\langle\frac{\tilde{\boldsymbol{A}}\boldsymbol{c}}{\|\tilde{\boldsymbol{A}}\boldsymbol{c}\|_2},\mathrm{vec}(\mathcal{A}-\Theta)\right\rangle.\end{aligned}\tag{15}$$

Therefore, we have

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\mathcal{C}_1,\mathcal{C}_2\in([0,1]^k)^{\otimes m}}\left\langle\frac{\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))}{\|\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))\|_F},\mathcal{A}-\Theta\right\rangle\geq t\right)\\ &\leq\mathbb{P}\left(\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\boldsymbol{c}\in\mathbb{R}^{2k^m}}\left\langle\frac{\tilde{\boldsymbol{A}}\boldsymbol{c}}{\|\tilde{\boldsymbol{A}}\boldsymbol{c}\|_2},\mathrm{vec}(\mathcal{A}-\Theta)\right\rangle\geq t\right)\\ &\leq\sum_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\mathbb{P}\left(\left\langle\frac{\tilde{\boldsymbol{A}}\boldsymbol{c}}{\|\tilde{\boldsymbol{A}}\boldsymbol{c}\|_2},\mathrm{vec}(\mathcal{A}-\Theta)\right\rangle\geq t\right)\\ &\leq\sum_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\mathbb{P}\left(\max_{\boldsymbol{c}\in\mathbb{R}^{\mathrm{rank}(\tilde{\boldsymbol{A}})}}\left\langle\frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2},\boldsymbol{x}\right\rangle\geq t\right),\end{aligned}\tag{16}$$

where $\boldsymbol{x}\in\mathbb{R}^{\mathrm{rank}(\tilde{\boldsymbol{A}})}$ is a sub-Gaussian vector with proxy variance $\sigma^2 = 1/4$ from Lemma 6. We complete the

proof from bounding the last probability in (16),

$$\begin{aligned}\sum_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}} \mathbb{P}\left(\max_{\boldsymbol{c}\in\mathbb{R}^{\mathrm{rank}(\boldsymbol{A})}} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2}, \boldsymbol{x}\right\rangle \geq t\right) &\leq \sum_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}} \exp\left(-C_1t^2 + C_2\mathrm{rank}(\tilde{\boldsymbol{A}})\right)\\ &\leq |\mathcal{M}|^2\exp(-C_1t^2+C_2k^m)\\ &= \exp(-C_1t^2+C_2k^m+2n\log k),\end{aligned}$$

where the first inequality is from [Phillippe Rigollet, 2015, Theorem 1.19] and the second inequality is from the fact $\mathrm{rank}(\tilde{\boldsymbol{A}}) \leq 2k^m$. □

Lemma 8 (Sub-Gaussian moment bounds)**.** The following holds for some constant $C_1, C_2 > 0$.

1. The first moment:

$$\mathbb{E}_{\mathcal{A}}\left[\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\mathcal{C}_1,\mathcal{C}_2\in([0,1]^k)^{\otimes m}}\left\langle\frac{\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))}{\|\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))\|_F},\mathcal{A}-\Theta\right\rangle\middle|\boldsymbol{\xi}\right] \leq C_1\sqrt{n\log k+k^m}$$

2. The second moment:

$$\mathbb{E}_{\mathcal{A}}\left[\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\mathcal{C}_1,\mathcal{C}_2\in([0,1]^k)^{\otimes m}}\left\langle\frac{\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))}{\|\mathrm{cut}(\Theta(\boldsymbol{M}_1,\mathcal{C}_1))-\mathrm{cut}(\Theta(\boldsymbol{M}_2,\mathcal{C}_2))\|_F},\mathcal{A}-\Theta\right\rangle^2\middle|\boldsymbol{\xi}\right] \leq C_2\left(n\log k+k^m\right)$$

Proof. First, we prove the first moment bound. From (15), it suffices to bound

$$\mathbb{E}_{\mathcal{A}}\left[\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\boldsymbol{c}\in\mathbb{R}^{2k^m}}\left\langle\frac{\tilde{\boldsymbol{A}}\boldsymbol{c}}{\|\tilde{\boldsymbol{A}}\boldsymbol{c}\|_2},\mathrm{vec}(\mathcal{A}-\Theta)\right\rangle\middle|\boldsymbol{\xi}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\boldsymbol{c}\in\mathbb{R}^{\mathrm{rank}(\tilde{\boldsymbol{A}})}}\left\langle\frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2},\boldsymbol{x}\right\rangle\middle|\boldsymbol{\xi}\right], \tag{17}$$

where $\boldsymbol{x}\in\mathbb{R}^{\mathrm{rank}(\tilde{\boldsymbol{A}})}$ is a sub-Gaussin vector with proxy variance $\sigma^2 = 1/4$ from Lemma 6. Notice that $\tilde{\boldsymbol{A}}$ is a function of $\boldsymbol{M}_1, \boldsymbol{M}_2$. Let $\mathcal{N}$ be a 1/2-net of a unit ball $\mathcal{B}_2 = \left\{\frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2} : \boldsymbol{c}\in\mathbb{R}^{\mathrm{rank}(\tilde{\boldsymbol{A}})}\right\}$ such that $|\mathcal{N}| \leq C^{\mathrm{rank}(\tilde{\boldsymbol{A}})}$ for constant $C > 0$. We can find such 1/2-net because $\mathcal{B}_2$ is in $\mathrm{rank}(\tilde{\boldsymbol{A}})$-dimension [Phillippe Rigollet, 2015, Lemma 1.18]. Then for any $\boldsymbol{c}_1\in\mathcal{B}_2$, there exists $\boldsymbol{c}_2\in\mathcal{N}$ such that $\|\boldsymbol{c}_1-\boldsymbol{c}_2\|_2 \leq 1/2$. For any $\boldsymbol{x}\in\mathbb{R}^{\mathrm{rank}(\tilde{\boldsymbol{A}})}$, we have

$$\begin{aligned}|\langle\boldsymbol{c}_1,\boldsymbol{x}\rangle| &\leq |\langle\boldsymbol{c}_1-\boldsymbol{c}_2,\boldsymbol{x}\rangle| + |\langle\boldsymbol{c}_2,\boldsymbol{x}\rangle|\\ &= \|\boldsymbol{c}_1-\boldsymbol{c}_2\|_2\left|\left\langle\frac{\boldsymbol{c}_1-\boldsymbol{c}_2}{\|\boldsymbol{c}_1-\boldsymbol{c}_2\|_2},\boldsymbol{x}\right\rangle\right| + |\langle\boldsymbol{c}_2,\boldsymbol{x}\rangle|\\ &\leq \frac{1}{2}\sup_{\boldsymbol{c}_1\in\mathcal{B}_2}|\langle\boldsymbol{c}_1,\boldsymbol{x}\rangle| + |\langle\boldsymbol{c}_2,\boldsymbol{x}\rangle|.\end{aligned}$$

Taking sup on both sides yields

$$\sup_{\boldsymbol{c}_1\in\mathcal{B}_2}|\langle\boldsymbol{c}_1,\boldsymbol{x}\rangle| \leq 2\sup_{\boldsymbol{c}_2\in\mathcal{N}}|\langle\boldsymbol{c}_2,\boldsymbol{x}\rangle|.$$

From the inequality, (17) becomes

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\left[\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\boldsymbol{c}\in\mathbb{R}^{\mathrm{rank}(\tilde{\boldsymbol{A}})}}\left\langle\frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2},\boldsymbol{x}\right\rangle\middle|\boldsymbol{\xi}\right] &= \mathbb{E}_{\boldsymbol{x}}\left[\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\boldsymbol{c}_1\in\mathcal{B}_2}\langle\boldsymbol{c}_1,\boldsymbol{x}\rangle\middle|\boldsymbol{\xi}\right]\\ &\leq 2\mathbb{E}_{\boldsymbol{x}}\left[\sup_{\boldsymbol{M}_1,\boldsymbol{M}_2\in\mathcal{M}}\sup_{\boldsymbol{c}_2\in\mathcal{N}}\langle\boldsymbol{c}_2,\boldsymbol{x}\rangle\middle|\boldsymbol{\xi}\right].\end{aligned}$$

Notice the number of possible combination of $(\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{c}_2)$ is $|\mathcal{M}|^2 C^{\mathrm{rank}(\tilde{\boldsymbol{A}})} \le k^{2n} C^{2k^m}$. Therefore, [Phillippe Rigollet, 2015, Theorem 1.14] gives us the first moment bound.

For the second moment bound, it suffices to show

$$\mathbb{E}_{\boldsymbol{x}} \left[\sup_{\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathcal{M}} \sup_{\boldsymbol{c}_2 \in \mathcal{N}} \langle \boldsymbol{c}_2, \boldsymbol{x} \rangle^2 \middle| \boldsymbol{\xi} \right] \le C \left(n \log k + k^m \right),$$

for some constant $C > 0$. Since the number of possible combination of $(\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{c}_2)$ is less than $k^{2n} C^{2k^m}$, and $\langle \boldsymbol{c}_2, \boldsymbol{x} \rangle$ is again sub-Gaussian with proxy variance $\sigma^2 = 1/4$, it is enough to show that for independent sub-Gausssian random variables $\{X_i\}_{i=1}^N$ with proxy variance σ^2,

$$\mathbb{E} \left[\max_{1 \le i \le N} X_i^2 \right] \le 8\sigma^2 \log N. \tag{18}$$

Then, plugging $N = k^2 n C^{2k^m}$ and $\sigma^2 = 1/4$ into (18) completes the proof. (18) is proved as follows

$$\begin{aligned}
\mathbb{E} \left[\max_{1 \le i \le N} X_i^2 \right] &= \frac{1}{s} \mathbb{E} \left[\log e^{s \max_{1 \le i \le N} X_i^2} \right] \\
&\le \frac{1}{s} \log \mathbb{E} \left[e^{s \max_{1 \le i \le N} X_i^2} \right] && \text{[Jensen inequality]} \\
&\le \frac{1}{s} \log \sum_{1 \le i \le N} \mathbb{E} \left[e^{s X_i^2} \right] \\
&\le \frac{1}{s} \log \sum_{1 \le i \le N} \left(1 + \sum_{k=2}^{\infty} \frac{s^k \mathbb{E} X_i^{2k}}{k!} \right) \\
&\le \frac{1}{s} \log \sum_{1 \le i \le N} \left(1 + 2 \sum_{k=2}^{\infty} (2s\sigma^2)^k \right) && \text{[Phillippe Rigollet, 2015, Lemma 1.4]} \\
&= \frac{1}{s} \log \sum_{1 \le i \le N} \left(1 + 8s^2\sigma^4 \sum_{k=0}^{\infty} (2s\sigma^2)^k \right) \\
&\le \frac{1}{s} \log \sum_{1 \le i \le N} \left(1 + 16 s^2 \sigma^4 \right)
\end{aligned}$$

Taking $s = 1/(4\sigma^2)$ yields,

$$\mathbb{E} \left[\max_{1 \le i \le N} X_i^2 \right] \le 4\sigma^2 (\log N + \log 2) \le 8\sigma^2 \log N.$$

□

References

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

Jan-Christian Hitter Phillippe Rigollet. High dimensional statistics. *Lecture notes for course 18S997*, 2015.