Hypergraphon estimation error updated

Chanwoo Lee

March 10, 2021

1 Notation and problem setting

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

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

We denote an index of m-uniform hyperedges as $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_m) \in [n]^m$ and a membership vector of m-vertices as $\boldsymbol{a} = (a_1,\ldots,a_m) \in [k]^m$. 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

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

where $f\colon [0,1]^m \to [0,1]$ is a symmetric function called hypergraphon such that $f(\xi_{\omega_1},\ldots,\xi_{\omega_m}) = f(\xi_{\sigma(\omega_1)},\ldots,\xi_{\sigma(\omega_m)})$ for all permutation $\sigma\colon [n] \to [n]$. Conventionally, we set $\Theta_{\boldsymbol{\omega}} = 0$ for all $\boldsymbol{\omega} \in [n]^m \setminus E$. $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_n)$ is a random vector of unobserved (latent) random variables i.i.d. drawn from uniform distribution U$[0,1]$. We call Θ the probability tensor.

2 Probability tensor estimation

In this section, we estimate the probability tensor in (1). Let $\tilde{\Theta}$ be a minimizer of the least square error from the adjacency tensor $\mathcal{A}$,

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

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 $\hat{\Theta} = \text{cut}(\tilde{\Theta})$ such that

$$\text{cut}(\Theta_{\boldsymbol{\omega}}) = \begin{cases} \Theta_{\boldsymbol{\omega}} & \text{if } \boldsymbol{\omega} \in E, \\ 0 & \text{if } \boldsymbol{\omega} \in [n]^m \setminus E. \end{cases} \tag{2}$$

Notice $\|\mathcal{A}-\hat{\Theta}\|_F^2 \leq \|\mathcal{A}-\Theta\|_F^2$ for any k block tensor $\Theta \in \text{cut}(\mathcal{P}_k)$. We construct estimation error bounds under the two cases when hypergraphon is k-piecewise constant and Hölder continuous.

2.1 Under k-piecewise constant hypergraphon model (Stochastic block model)

Here, we consider the case when the probability tensor is from k-piecewise constant hypergraphon.

Definition 1. For an integer $k < n$, a function $f\colon [0,1]^m \to [0,1]$ is k-piecewise constant, denoted as $f \in \mathcal{P}(k)$, if there exist $\mathcal{C} \in ([0,1]^k)^{\otimes m}$ and $\phi\colon [0,1] \to [k]$ such that

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

Notice that if $f \in \mathcal{P}(k)$, our model is equivalent to the stochastic block model, $\Theta \in \text{cut}(\mathcal{P}_k)$.

Theorem 2.1 (Stochastic block model)**.** *Let $\hat{\Theta}$ be the estimator from (2). Suppose true probability tensor $\Theta \in \text{cut}(\mathcal{P}_k)$ for fixed block size k. Then, there exists two constants $C_1, C_2 > 0$, such that*

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

with probability at least $1-\exp\left(-C_2(n\log k + k^m)\right)$. Furthermore, expected mean square error is bounded by

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

for some constant $C > 0$.

Remark 1. 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 2. 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 this clustering uncertainty is reflected by logarithm in the clustering rate.

2.2 Under Hölder continuous hypergraphon model

In this subsection, we assume that a hypergraphon f is α-Hölder continuous with a constant L.

Definition 2. 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

$$|f(x) - f(y)| \leq L\|x-y\|_\alpha^\alpha \quad \text{for all } x, y \in [0,1]^m, \tag{3}$$

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

The following lemma shows that the probability tensor generated from Hölder continuous hypergraphon is approximated well by a block probability tensor.

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

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

Combining Lemma 1 and Theorem 2.1 yields the upper bound of estimation error for the probability tensor from Hölder continuous hypergraphon.

Theorem 2.2 (Hölder continuous 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 (2). Then, there exist two constants $C_1, C_2 > 0$ such that,

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

with probability at least $1 - \exp\left(-C_2\left(n\log n + n^{\frac{m^2}{m+2\alpha}}\right)\right)$ uniformly over $f \in \mathcal{H}(\alpha, L)$. Furthermore, the expected mean square error is bounded by

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

for some constant $C > 0$.

Remark 3. 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 explanation 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 in n-vertices remains roughly the same (which contributes to the clustering rate) when m increases. Precisely, it increases by a factor $\frac{m}{m+2\alpha}$. That means that the hypergraphon estimation becomes harder due to the increased number of parameters while cluster estimation becomes relatively easier because we have roughly fixed combinations of clusters among n vertices but with more information about interaction among clusters. When $m = 2$, Theorem 2.2 reduces to the results in Gao et al. [2015].

Remark 4. 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 2.2 becomes

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

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.

3 Hypergraphon estimation

In this section, we estimate hypergraphon function. 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:

$$f_\Theta(x_1,\ldots,x_m) = \Theta_{\lfloor x_1 \rfloor,\ldots,\lfloor x_m \rfloor}.$$

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

$$\delta^2(\hat{f}, f) := \inf_{\tau\in\mathcal{T}} \int_{(0,1)^m} |f(\tau(\boldsymbol{x})) - \hat{f}(\boldsymbol{x})|^2 d\boldsymbol{x}, \tag{5}$$

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 as in (5).

We construct the upper bound of expected mean square error for hypergraphon based on the triangular inequality,

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

We have already found upper bound of (i) in previous section in two cases. We construct the upper bounds of expected mean square error by finding error bound of (ii) in two cases: piecewise constant hypergraphon and Hölder continuous hypergraphon.

3.1 k-piecewise constant hypergraphon

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

Lemma 2 (Agnostic error for k-piecewise constant hypergraphon)**.** Consider the hypergraphon model (1). For all integers $k \le n$, $f \in \mathcal{F}(k)$, we have

$$\mathbb{E}\left[\delta^2(f_\Theta, f)\right] \le Cm\sqrt{\frac{k}{n}},$$

for some constant $C > 0$.

By Theorem 2.1 and Lemma 2, we have the following theorem.

Theorem 3.1 (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 (2). Then, there exists a positive constant $C > 0$ only depending on m and L such that

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

Remark 5. The error rate is dominated by agnostic error $\sqrt{k/n}$ for all combinations of $\{(m,\alpha)\}_{m\ge 2,\alpha\in(0,1]}$. 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 remains roughly 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 Hölder continuous hypergraphon

The egonostic error associated to α-Hölder continuous hypergraphon is bounded as follows.

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

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

for some universal constant $C > 0$.

By Theorem 2.2 and Lemma 3, we have the following theorem.

Theorem 3.2 (Mean square error of 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 (2). Then, there exists a positive constant $C > 0$ only depending on m and L such that

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

Remark 6. The error rate is dominated by agnostic error $n^{-\alpha}$ for all combinations of $\{(m,\alpha)\}_{m\geq 2, \alpha\in(0,1]}$ except $(m,\alpha) = (2,1)$, when $\log n/n$ is the dominating error term. This phenomenon can be explained in the same way in Remark 5

4 Proofs

4.1 Proof of Theorem 2.1

Proof. First, we prove the probability tail bound. From the definition of $\hat{\Theta}$, we have

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

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

$$\begin{aligned}\|\hat{\Theta} - \Theta^*\|_F^2 &\leq 2\langle \hat{\Theta} - \Theta, \mathcal{A} - \Theta\rangle \qquad (7)\\ &= 2\|\hat{\Theta} - \Theta\|_F \left\langle \frac{\hat{\Theta} - \Theta}{\|\hat{\Theta} - \Theta\|_F}, \mathcal{A} - \Theta\right\rangle \\ &\leq 2\|\hat{\Theta} - \Theta\|_F \sup_{\Theta'\in\mathcal{P}_k}\sup_{\Theta''\in\mathcal{P}_k}\left\langle \frac{\text{cut}(\Theta') - \text{cut}(\Theta'')}{\|\text{cut}(\Theta') - \text{cut}(\Theta'')\|_F}, \mathcal{A} - \Theta\right\rangle.\end{aligned}$$

Notice the last equality holds because true probability tensor has a block structure i.e. $\Theta \in \text{cut}(\mathcal{P}_k)$. Let $\mathcal{M} = \{\boldsymbol{M} \colon \boldsymbol{M}$ is the collection of membership matrices$\}$. Then, we have

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

Notice that $\mathcal{A}-\Theta$ is sub-Gaussian with proxy parameter $\sigma^2 = 1/4$. By Lemma 5, we have

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

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

For the expectation bound, let $\epsilon^2 = C_1\left(\left(\frac{k}{n}\right)^m + \frac{\log k}{n^{m-1}}\right)$. Then we have

$$\begin{aligned}
\frac{1}{n^m}\mathbb{E}\|\hat{\Theta}-\Theta\|_F^2 &= \mathbb{E}\left(n^{-m}\|\hat{\Theta}-\Theta\|_F^2\mathbb{1}_{\{n^{-m}\|\hat{\Theta}-\Theta\|_F^2\geq\epsilon^2\}}\right) + \mathbb{E}\left(n^{-m}\|\hat{\Theta}-\Theta\|_F^2\mathbb{1}_{\{n^{-m}\|\hat{\Theta}-\Theta\|_F^2<\epsilon^2\}}\right)\\
&\leq \mathbb{P}\left(n^{-m}\|\hat{\Theta}-\Theta\|_F^2 \geq \epsilon^2\right) + \epsilon^2\\
&\leq \exp\left(-C_2(n\log k + k^m)\right) + C_1\left(\left(\frac{k}{n}\right)^m + \frac{\log k}{n^{m-1}}\right),
\end{aligned} \tag{8}$$

where the first inequality is from the fact that $n^{-m}\|\hat{\Theta}-\Theta\|_F^2 \leq 1$ and the second inequality from the tail bound we have constructed above. Since $\left(\frac{k}{n}\right)^m + \frac{\log k}{n^{m-1}}$ dominates the term in (8), the proof is completed.

□

4.2 Proof of Lemma 1

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

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

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

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

□

4.3 Proof of Theorem 2.2

Proof. First, we prove the probability tail bound. By Lemma 1, we can always find a block tensor Θ^* close to the true probability tensor Θ such that

$$\frac{1}{n^m}\|\Theta^*-\Theta\|_F^2 \leq m^2L^2\left(\frac{1}{k^2}\right)^\alpha. \tag{9}$$

By triangular inequality,

$$\|\hat{\Theta}-\Theta\|_F^2 \leq 2\underbrace{\|\hat{\Theta}-\Theta^*\|_F^2}_{\text{(i)}} + 2\underbrace{\|\Theta^*-\Theta\|_F^2}_{\text{(ii)}}. \tag{10}$$

Since we have the error bound (ii) as in (9), we find the upper bound of the error (i). Based on definition of $\hat{\Theta}$, we have the following inequality similar to (7).

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

It suffices to bound the inner product term because of (9). Notice Θ^* is a block tensor such that $\Theta^* \in \text{cut}(\mathcal{P}(k))$. Therefore, by the same way in the proof of Theorem 2.1, we obtain

$$\mathbb{P}\left(\left\langle \frac{\hat{\Theta}-\Theta^*}{\|\hat{\Theta}-\Theta^*\|_F}, \mathcal{A}-\Theta\right\rangle \geq t\right) \leq \exp(-C_1t^2 + C_2k^m + 2n\log k),$$

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

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

with probability at least $1-\exp\left(-C_2(n\log k + k^m)\right)$. Combinations of two error bounds in (10) and setting $k = \lceil n^{\frac{m}{m+2\alpha}}\rceil$, completes the theorem.

The moment bound follows by the same argument in (8). □

4.4 Proof of Lemma 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',

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

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

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

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

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

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

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

Finally, we have

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

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

4.5 Proof of Lemma 3

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 $\Theta'_{\boldsymbol{\omega}} = f(\xi_{\omega_1}, \dots, \xi_{\omega_m})$ 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',

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

for a universal constant $C > 0$. This is because $|f_\Theta(\boldsymbol{x}) - f_{\Theta'}(\boldsymbol{x})| = 0$ outside of a set of measure $\left(n^m - \binom{n}{m} n!\right)/n^m \leq Cm^2/n$. Hence it suffices to prove that

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

for some constant $C > 0$. We have

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

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

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

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

$$|f(x_1, \dots, x_m) - f(\xi_1, \dots, \xi_m)|^2 \leq 2\left| f(x_1, \dots, x_m) - f\left(\frac{\sigma(i_1)}{n+1}, \dots \frac{\sigma(i_m)}{n+1}\right)\right|^2 \qquad (11)$$

$$+ 2\left|f\left(\frac{\sigma(i_1)}{n+1},\dots\frac{\sigma(i_m)}{n+1}\right) - f(\xi_{(\sigma(i_1))},\dots,\xi_{(\sigma(i_m))})\right|^2$$
$$+ 2\left|f(\xi_{(\sigma(i_1))},\dots,\xi_{(\sigma(i_m))}) - f(\xi_{i_1},\dots,\xi_{i_m})\right|^2,$$

where $\xi_{(\ell)}$ denotes the ℓ-th largest element of the set $\{\xi_1,\dots,\xi_n\}$. 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 (11) is 0 almost surely.

For the first summand in (11), notice $(\sigma(i_1)/(n+1),\dots\sigma(i_m)/(n+1)) \in \prod_{\ell=1}^m[(\sigma(i_\ell)-1)/n, \sigma(i_\ell)/n]$. From (3), we obtain

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

Integrating and taking expectation on the first summand yields,

$$\mathbb{E}\left[\sum_{i_1,\dots,i_m=1}^n \int_{(\sigma(i_1)-1)/n}^{\sigma(i_1)/n}\cdots\int_{(\sigma(i_m)-1)/n}^{\sigma(i_m)/n}\left|f(x_1,\dots,x_m) - f\left(\frac{\sigma(i_1)}{n+1},\dots\frac{\sigma(i_m)}{n+1}\right)\right|^2 dx_1\cdots dx_m\right] \le m^2L^2\left(\frac{1}{n}\right)^{2\alpha}. \tag{12}$$

With (3), the second summand on (11) is bounded,

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

Integrating and taking expectation on the second summand yields,

$$\begin{aligned}&\mathbb{E}\left[\sum_{i_1,\dots,i_m=1}^n \int_{(\sigma(i_1)-1)/n}^{\sigma(i_1)/n}\cdots\int_{(\sigma(i_m)-1)/n}^{\sigma(i_m)/n}\left|f\left(\frac{\sigma(i_1)}{n+1},\dots\frac{\sigma(i_m)}{n+1}\right) - f(\xi_{(\sigma(i_1))},\dots,\xi_{(\sigma(i_m))})\right|^2 dx_1\cdots dx_m\right]\\
&\le \left[\sum_{i_1,\dots,i_m=1}^n \int_{(\sigma(i_1)-1)/n}^{\sigma(i_1)/n}\cdots\int_{(\sigma(i_m)-1)/n}^{\sigma(i_m)/n} 2L^2\sum_{\ell=1}^m\left|\frac{\sigma(i_\ell)}{n+1} - \xi_{(\sigma(i_\ell))}\right|^{2\alpha} dx_1\cdots dx_m\right]\\
&= 2mL^2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left|\frac{i}{n+1} - \xi_{(i)}\right|^{2\alpha}\right]\\
&\le 2mL^2\max_{i=1,\dots,n}\mathbb{E}\left[\left|\frac{i}{n+1} - \xi_{(i)}\right|^{2\alpha}\right]\\
&\le 2mL^2\max_{i=1\dots,n}\left[\mathrm{Var}(\xi_{(i)})\right]^\alpha \le CmL^2\left(\frac{1}{n}\right)^\alpha,\end{aligned} \tag{13}$$

where we have used $\mathbb{E}(\xi_{(\ell)}) = \ell/(n+1)$, $\mathrm{Var}(\xi_{(\ell)}) \le C/n$ and Jensen's inequality. Combining (12) and (13) proves the lemma. □

5 Technical lemmas

Lemma 4 (Gaussian maxima under full embedding)**.** Let $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$ be a deterministic matrix with rank $r \le \min(d_1, d_2)$. Let $\boldsymbol{y} \in \mathbb{R}^{d_2}$ 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

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

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

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

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

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

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 7. 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

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

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 5 (Sub-Gaussian tail bound)**.**

$$\mathbb{P}\left(\sup_{\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathcal{M}} \sup_{\mathcal{C}_1, \mathcal{C}_2 \in ([0,1]^k)^{\otimes m}} \left\langle \frac{\text{cut}(\Theta(\boldsymbol{M}_1, \mathcal{C}_1)) - \text{cut}(\Theta(\boldsymbol{M}_2, \mathcal{C}_2))}{\|\text{cut}(\Theta(\boldsymbol{M}_1, \mathcal{C}_1)) - \text{cut}(\Theta(\boldsymbol{M}_2, \mathcal{C}_2))\|_F}, \mathcal{A} - \Theta \right\rangle \geq t \right) \leq \exp(-C_1 t^2 + C_2 k^m + 2n \log k),$$

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 $\Theta_{\boldsymbol{\omega}} = \text{vec}(\Theta)_{\omega'}$. Notice that

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

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

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

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

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

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathcal{M}} \sup_{\mathcal{C}_1, \mathcal{C}_2 \in ([0,1]^k)^{\otimes m}} \left\langle \frac{\mathrm{cut}(\Theta(\boldsymbol{M}_1, \mathcal{C}_1)) - \mathrm{cut}(\Theta(\boldsymbol{M}_2, \mathcal{C}_2))}{\|\mathrm{cut}(\Theta(\boldsymbol{M}_1, \mathcal{C}_1)) - \mathrm{cut}(\Theta(\boldsymbol{M}_2, \mathcal{C}_2))\|_F}, \mathcal{A} - \Theta \right\rangle \geq t \right) \\
&\quad \leq \mathbb{P}\left(\sup_{\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathcal{M}} \sup_{\boldsymbol{c} \in \mathbb{R}^{2k^m}} \left\langle \frac{\tilde{\boldsymbol{A}}\boldsymbol{c}}{\|\tilde{\boldsymbol{A}}\boldsymbol{c}\|_2}, \mathrm{vec}(\mathcal{A} - \Theta) \right\rangle \geq t \right) \\
&\quad \leq \sum_{\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathcal{M}} \mathbb{P}\left(\left\langle \frac{\tilde{\boldsymbol{A}}\boldsymbol{c}}{\|\tilde{\boldsymbol{A}}\boldsymbol{c}\|_2}, \mathrm{vec}(\mathcal{A} - \Theta) \right\rangle \geq t \right) \\
&\quad \leq \sum_{\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathcal{M}} \mathbb{P}\left(\max_{\boldsymbol{c} \in \mathbb{R}^{\mathrm{rank}(\tilde{\boldsymbol{A}})}} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2}, \boldsymbol{x} \right\rangle \geq t \right), \qquad (14)
\end{aligned}$$

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 4. We complete the proof from bounding the last probability in (14),

$$\begin{aligned}
\sum_{\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathcal{M}} \mathbb{P}\left(\max_{\boldsymbol{c} \in \mathbb{R}^{\mathrm{rank}(\boldsymbol{A})}} \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|_2}, \boldsymbol{x} \right\rangle \geq t \right) &\leq \sum_{\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathcal{M}} \exp\left(-C_1 t^2 + C_2 \mathrm{rank}(\tilde{\boldsymbol{A}}) \right) \\
&\leq |\mathcal{M}|^2 \exp(-C_1 t^2 + C_2 k^m) \\
&= \exp(-C_1 t^2 + C_2 k^m + 2n \log k),
\end{aligned}$$

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

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.