

Extension to sparse regime and algorithm performance in smooth settings

Chanwoo Lee
April 12, 2021

1 Algorithm performance in smooth settings

We assume that $\mathcal{A}_\omega \sim \text{Bernoulli}(\Theta_\omega)$, where

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

where $f: [0, 1]^m \rightarrow [0, 1]$ is a symmetric function called hypergraphon such that $f(\xi_{\omega_1}, \dots, \xi_{\omega_m}) = f(\xi_{\sigma(\omega_1)}, \dots, \xi_{\sigma(\omega_m)})$ for all permutation $\sigma: [m] \rightarrow [m]$. I checked our algorithm performance under this hypergraphon model with different symmetric function $f: [0, 1]^3 \rightarrow [0, 1]$.

- **Smooth 1:** $f(x_1, x_2, x_3) = 1/(1 + \exp(-(x_1^2 + x_2^2 + x_3^2)))$.
- **Smooth 2:** $f(x_1, x_2, x_3) = x_1 x_2 x_3$.
- **Smooth 3:** $f(x_1, x_2, x_3) = \log(1 + \max(x_1, x_2, x_3))$.
- **Smooth 4:** $f(x_1, x_2, x_3) = \exp(-\min(x_1, x_2, x_3))$.

Figure 1 shows the distributions of Θ from each model when $n = 100$.

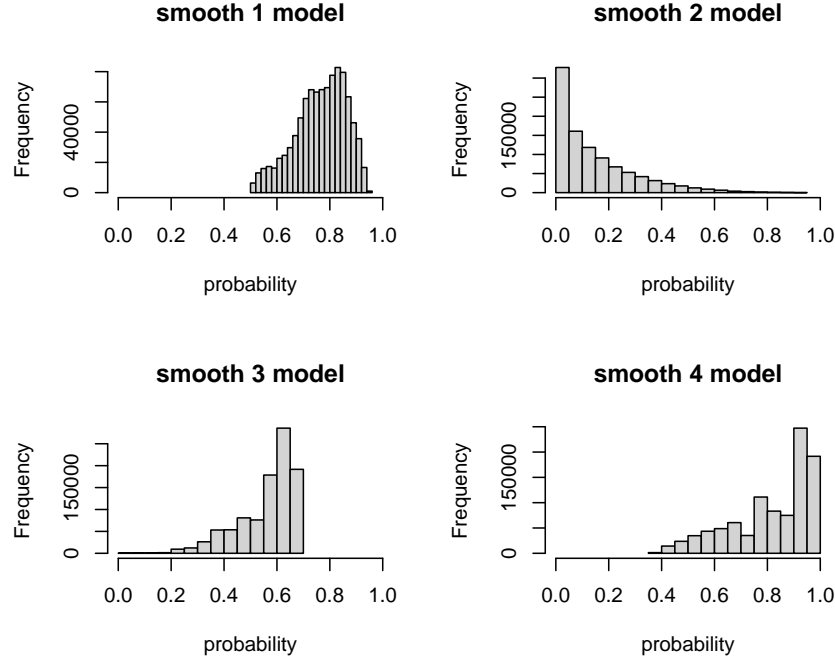


Figure 1: Empirical density of the probability tensor Θ for each smooth model (Smooth 1-4).

In hypergraphon model, there is no clusters. So I choose to use $k \in \{5, 10, 15, 20\}$ for each $n \in \{50, 100, 150, 200, 250\}$. I use `functions_sbm` for the updates. Figure 2 shows the MSE result according to different smooth models. It turns out that our algorithm works great in the smooth settings except small group size case in Smooth 3 and Smooth 4. I found that `functions_sbm` is sometimes trapped in local minimums from which

`functions_sbm2` escape. Figure 3 shows that `functions_sbm2` succeeded to find the good optimal points where `functions_sbm` failed.

Figure 4 shows the optimal group size for each smooth model when the number of mode is fixed to 100.

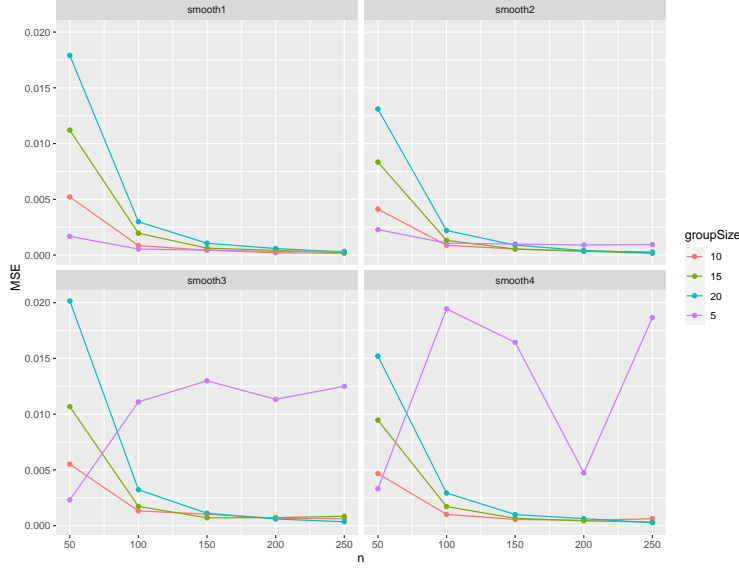


Figure 2: MSE error depending on smooth model, the number of node $n \in \{50, 100, 150, 200, 250\}$ and the number of clusters $k \in \{5, 10, 15, 20\}$. `functions_sbm` is used for the algorithm. `functions_sbm` is used on red lines (sbm) while `functions_sbm2` on blue lines (sbm2).

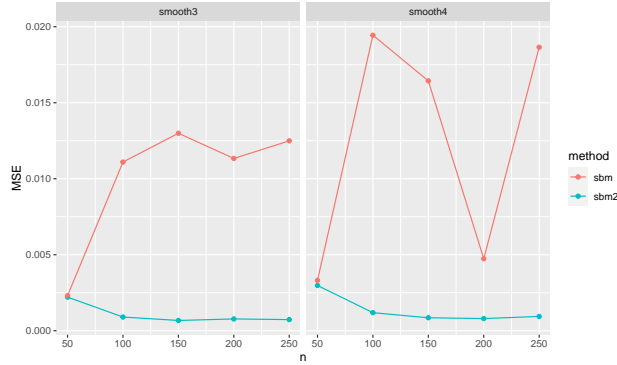


Figure 3: MSE errors in Smooth 3 and 4 depending on the number of nodes. The cluster size is set to be 5 where previous `functions_sbm` failed to escape local minimums (Figure 2).

2 Extension to sparse regime

We denote $\rho \in [0, 1]$ as the sampling probability or the sparsity parameter. When the observed tensor $\mathcal{A} \in \{0, 1\}^{d_1 \times \dots \times d_K}$ is complete, we interpret ρ as the sparsity parameter. On the other hand, we interpret ρ as the sampling probability when the data is incomplete, i.e.,

$$\mathbb{P}[\mathcal{A}_\omega \text{ is observed} | \Theta_\omega^{\text{true}}] = \rho.$$

If we assign missing entries as 0 in \mathcal{A} , the marginal probability of observed network being connected has

$$\mathbb{P}(\mathcal{A}_\omega = 1) = \rho \Theta_\omega^{\text{true}},$$

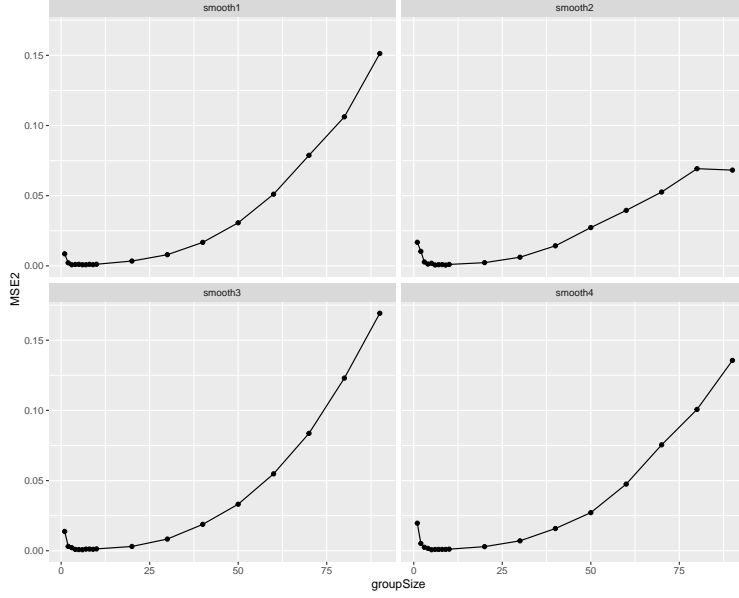


Figure 4: MSE errors for different group size fixing $n = 100$ depending on different smooth model 1-4.

for all $\rho \in E$ in both complete and incomplete cases. Here we choose to call the parameter ρ the sparsity parameter. Notice that when $\rho = 1$, the problem reduces to previous settings.

From known ρ , we estimate Θ^{true} by

$$\hat{\Theta} = \text{cut}(\tilde{\Theta}), \quad \text{where } \tilde{\Theta} = \arg \min_{\Theta \in \mathcal{P}_k} \sum_{\omega \in E} |\mathcal{A}_\omega - \rho \Theta_\omega|^2. \quad (1)$$

With adaptation of the new parameter ρ , we modify previous theorems incorporating ρ .

2.1 Probability tensor estimation

Under k -piecewise constant hypergraphon model, we construct theoretical guarantees for estimated probability tensor from observed network tensor $\mathcal{A} \in \{0, 1\}^{d_1 \times \dots \times d_m}$.

Theorem 2.1 (Stochastic block model with the sparsity parameter ρ). Let $\hat{\Theta}$ be the estimator from (1). Suppose true probability tensor $\Theta^{\text{true}} \in \text{cut}(\mathcal{P}_k)$ for fixed block size k . Then, there exists two constants $C_1, C_2 > 0$, such that

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

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

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

for some constant $C > 0$.

Remark 1. The sparsity parameter ρ makes both nonparametric and clustering rates worse by the same rate. For the nonparametric rate, the number of observation becomes $\mathcal{O}(\rho n^m)$ while the number of parameters remains the same as $\mathcal{O}(k^m)$. This reduced observation is reflected on the rate. Similarly, the number of possible k -clusters of n -vertices remains $\mathcal{O}(k^n)$ while the number of observation changes to $\mathcal{O}(\rho n^m)$.

Therefore, the clustering rate is also reduced by ρ .

Now we assume that a hypergraphon f is α -Hölder continuous with a constant L . We define a block average on a set $E_{z^{-1}(a)}$ for a given membership function z , a membership vector a , and a tensor $\Theta \in ([n])^{\otimes m}$ as

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

Notice that we define $\Theta_\omega = f(\xi_{\omega_1}, \dots, \xi_{\omega_m})$ for all $\omega = (\omega_1, \dots, \omega_m) \in E$ where there is no sparsity parameter ρ . Therefore the following block approximation lemma remains the same.

Lemma 2.1 (Block approximation). Suppose the true parameter Θ^{true} admits the hypergraphon model with $f \in \mathcal{H}(\alpha, L)$. For every integer $k \leq n$, there exists $z^*: [n] \rightarrow [k]$, satisfying

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

Theorem 2.2 (Hölder continuous hypergraphon model with the sparsity parameter ρ). Suppose the true parameter Θ^{true} admits the hypergraphon model with $f \in \mathcal{H}(\alpha, L)$. Let $\hat{\Theta}$ be the estimator from (1). Then, there exist two constants $C_1, C_2 > 0$ such that,

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

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

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

for some constant $C > 0$.

Remark 2. Notice that the recovery of the true probability tensor Θ^{true} is influenced by the expected number of observations, ρn^m . The probability $1 - \exp(-C_2 (n \log(\rho n^m) + (\rho n^m)^{\frac{m}{m+2\alpha}}))$ tells us that the sparsity parameter should be greater than $1/n^m$ to expect the meaningful MSE. Depending on constants m, α and the sparsity parameter ρ , convergence rate becomes

$$(\rho n^m)^{\frac{-2\alpha}{m+2\alpha}} + \frac{\log n}{\rho n^{m-1}} \asymp \begin{cases} (\rho n^m)^{\frac{-2\alpha}{m+2\alpha}} & \text{if } \rho \geq n^{1-m+2\alpha/m}, \\ \frac{\log n}{\rho n^{m-1}} & \text{if } \rho < n^{1-m+2\alpha/m}, \end{cases}$$

upto $\log n$ factors. The nonparametric rate tends to dominate the error when dense observation while the clustering rate dominates the error under the sparse regime.

3 Proof

3.1 Proof of Theorem 2.1

Proof. We consider two exclusive cases

1. Case 1: when $\|\hat{\Theta} - \Theta^{\text{true}}\|_F \leq \sqrt{C(k^m + n \log k)}/\rho$, then we have directly (2).
2. Case 2: when $\|\hat{\Theta} - \Theta^{\text{true}}\|_F > \sqrt{C(k^m + n \log k)}/\rho$.

By the definition of $\hat{\Theta}$ in (1), we have

$$\begin{aligned}\|\rho\hat{\Theta} - \rho\Theta^{\text{true}}\|_F^2 &\leq 2\langle \rho\hat{\Theta} - \rho\Theta^{\text{true}}, \mathcal{A} - \rho\Theta^{\text{true}} \rangle \\ &= 2\|\rho\hat{\Theta} - \rho\Theta^{\text{true}}\|_F \left\langle \frac{\hat{\Theta} - \Theta^{\text{true}}}{\|\hat{\Theta} - \Theta^{\text{true}}\|_F}, \mathcal{A} - \rho\Theta^{\text{true}} \right\rangle.\end{aligned}\tag{3}$$

Then,

$$\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2 \leq 2 \left| \left\langle \frac{\hat{\Theta} - \Theta^{\text{true}}}{\|\hat{\Theta} - \Theta^{\text{true}}\|_F}, \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle \right| \leq 2 \sup_{\mathbf{M}, \mathbf{M}' \in \mathcal{M}} \sup_{\mathcal{C}, \mathcal{C}' \in ([0,1]^k)^{\otimes m}} \left| \left\langle \mathcal{T}(\mathbf{M}, \mathbf{M}', \mathcal{C}, \mathcal{C}'), \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle \right|,$$

where $\mathcal{T} = \frac{\text{cut}(\Theta(\mathbf{M}, \mathcal{C})) - \text{cut}(\Theta(\mathbf{M}', \mathcal{C}'))}{\|\text{cut}(\Theta(\mathbf{M}, \mathcal{C})) - \text{cut}(\Theta(\mathbf{M}', \mathcal{C}'))\|_F}$. Under the event $\|\hat{\Theta} - \Theta^{\text{true}}\|_F > \sqrt{C(k^m + n \log k)}/\rho$ for some constant C , we have

$$|\mathcal{T}_\omega| \leq \frac{1}{\|\hat{\Theta} - \Theta^{\text{true}}\|_F} \leq \sqrt{\frac{\rho}{C(k^m + n \log k)}}, \text{ for all } \omega \in [n]^{\otimes m}.$$

Therefore, combination of Lemma 3.1 and 3.2 yields

$$\begin{aligned}\mathbb{P} \left(\sup_{\mathbf{M}, \mathbf{M}' \in \mathcal{M}} \sup_{\mathcal{C}, \mathcal{C}' \in ([0,1]^k)^{\otimes m}} \left| \left\langle \mathcal{T}(\mathbf{M}, \mathbf{M}', \mathcal{C}, \mathcal{C}'), \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle \right| \geq t \right) \\ \leq \exp(C'(k^m + n \log k)) \cdot \exp \left(- \min \left(\frac{\rho t^2}{24}, \frac{t \sqrt{C \rho (k^m + n \log k)}}{4} \right) \right),\end{aligned}$$

for some constant $C' > 0$. Setting $t = \sqrt{C_1(k^m + n \log k)}/\rho$ for sufficiently large C_1 depending $C > 0$ completes the proof.

Expectation bound follows from the probability tail bound. \square

Lemma 3.1. [Gao et al., 2016, Lemma 13] Let $\{\mathcal{A}_\omega\}_{\omega \in E}$ be independent sub-Gaussian random variables with mean $\rho\Theta_\omega$ and proxy variance σ^2 , where $\Theta_\omega \in [-M, M]$, $\rho \in [0, 1]$, and E is an index set. Then, for $|\lambda| \leq \rho/(M \vee \sigma)$, we have

$$\mathbb{E} e^{\lambda \left(\frac{\mathcal{A}_\omega - \rho\Theta_\omega}{\rho} \right)} \leq 2e^{(M^2 + 2\sigma^2)\lambda^2/\rho}.$$

Moreover, for $\sum_{\omega \in E} c_\omega^2 = 1$,

$$\mathbb{P} \left\{ \left| \sum_{\omega \in E} c_\omega \left(\frac{\mathcal{A}_\omega - \rho\Theta_\omega}{\rho} \right) \right| \geq t \right\} \leq 4 \exp \left\{ - \min \left(\frac{\rho t^2}{4(M^2 + 2\sigma^2)}, \frac{\rho t}{2(M \vee \sigma) \|c\|_\infty} \right) \right\},$$

for any $t > 0$.

Lemma 3.2. Let $\mathcal{X} \in (\mathbb{R}^n)^{\otimes m}$ be a random tensor drawn from any distributions. Denote \mathcal{B}_2 as a unit ball in $(\mathbb{R}^n)^{\otimes m}$ with $\|\cdot\|_2$ distance. Suppose that

$$\mathbb{P}(|\langle \mathcal{T}, \mathcal{X} \rangle| \geq t) \leq \phi(t), \text{ for all } t > 0 \text{ and } \mathcal{T} \in \mathcal{B}_2,\tag{4}$$

for some function $\phi: \mathbb{R}_+ \rightarrow [0, 1]$. Define

$$\mathcal{B}_c = \left\{ \frac{\text{cut}(\Theta(\mathbf{M}, \mathcal{C})) - \text{cut}(\Theta(\mathbf{M}', \mathcal{C}'))}{\|\text{cut}(\Theta(\mathbf{M}, \mathcal{C})) - \text{cut}(\Theta(\mathbf{M}', \mathcal{C}'))\|_F} : \mathbf{M}, \mathbf{M}' \in \mathcal{M} \text{ and } \mathcal{C}, \mathcal{C}' \in ([0, 1]^k)^{\otimes m} \right\}$$

Then,

$$\mathbb{P} \left(\sup_{\mathcal{T} \in \mathcal{B}_c} |\langle \mathcal{T}, \mathcal{X} \rangle| \geq t \right) \leq \exp(Ck^m + 2n \log k) \phi(t/2),$$

for some constant $C > 0$.

Proof. Notice that

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

With careful allocation, we always find a matrix $\tilde{\mathbf{A}} \in \{0,1\}^{n^m \times 2k^m}$ such that,

$$\text{vec}(\text{cut}(\Theta(\mathbf{M}_1, \mathcal{C}_1)) - \text{cut}(\Theta(\mathbf{M}_2, \mathcal{C}_2))) = \tilde{\mathbf{A}} \mathbf{c}.$$

Notice

$$\begin{aligned} \mathbb{P} \left(\sup_{\mathcal{T} \in \mathcal{S}} |\langle \mathcal{T}, \mathcal{X} \rangle| \geq t \right) &= \mathbb{P} \left(\sup_{\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}} \sup_{\mathcal{C}_1, \mathcal{C}_2 \in ([0,1]^k)^{\otimes m}} \left\langle \frac{\text{cut}(\Theta(\mathbf{M}_1, \mathcal{C}_1)) - \text{cut}(\Theta(\mathbf{M}_2, \mathcal{C}_2))}{\|\text{cut}(\Theta(\mathbf{M}_1, \mathcal{C}_1)) - \text{cut}(\Theta(\mathbf{M}_2, \mathcal{C}_2))\|_F}, \mathcal{X} \right\rangle \geq t \right) \\ &\leq \mathbb{P} \left(\sup_{\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}} \sup_{\mathbf{c} \in \mathbb{R}^{2k^m}} \left\langle \frac{\tilde{\mathbf{A}} \mathbf{c}}{\|\tilde{\mathbf{A}} \mathbf{c}\|_2}, \text{vec}(\mathcal{X}) \right\rangle \geq t \right) \\ &\leq \sum_{\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}} \mathbb{P} \left(\left\langle \frac{\tilde{\mathbf{A}} \mathbf{c}}{\|\tilde{\mathbf{A}} \mathbf{c}\|_2}, \text{vec}(\mathcal{X}) \right\rangle \geq t \right). \end{aligned}$$

Let $r := \text{rank}(\tilde{\mathbf{A}}) \leq 2k^m$ be the rank of $\tilde{\mathbf{A}}$. Then we express $\tilde{\mathbf{A}} \mathbf{c} = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{c} = \sum_{i=1}^r (\lambda_i \mathbf{v}_i^T \mathbf{c}) \mathbf{u}_i$, where $\mathbf{u}_i, \mathbf{v}_i$ and λ_i are i -th singular vectors and singular value respectively. Defining $\boldsymbol{\alpha} = (\lambda_1 \mathbf{v}_1^T \mathbf{c}, \dots, \lambda_r \mathbf{v}_r^T \mathbf{c}) \in \mathbb{R}^r$, we have,

$$\begin{aligned} \sum_{\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}} \mathbb{P} \left(\left\langle \frac{\tilde{\mathbf{A}} \mathbf{c}}{\|\tilde{\mathbf{A}} \mathbf{c}\|_2}, \text{vec}(\mathcal{X}) \right\rangle \geq t \right) &= \sum_{\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}} \mathbb{P} \left(\left\langle \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|_2}, (\mathbf{u}_1, \dots, \mathbf{u}_r)^T \text{vec}(\mathcal{X}) \right\rangle \geq t \right) \\ &\leq \sum_{\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}} \mathbb{P} \left(\max_{\mathbf{c} \in \mathbb{R}^r} \left\langle \frac{\mathbf{c}}{\|\mathbf{c}\|_2}, \mathbf{x} \right\rangle \geq t \right) \\ &\leq \sum_{\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}} \mathbb{P} \left(\max_{\mathbf{c} \in \mathcal{S}} \langle \mathbf{c}, \mathbf{x} \rangle \geq t \right), \end{aligned} \tag{5}$$

where we define $\mathbf{x} := (\mathbf{u}_1, \dots, \mathbf{u}_r)^T \text{vec}(\mathcal{X}) \in \mathbb{R}^r$ and \mathcal{S} is a unit sphere in \mathbb{R}^r . Notice the upper bound (4) still holds for \mathbf{x} by the orthonormality of $(\mathbf{u}_1, \dots, \mathbf{u}_r)$. Let \mathcal{S}' be a $1/2$ -net of \mathcal{S} with respect to the Euclidean norm that satisfies $|\mathcal{S}'| \leq 6^r$. Observed that for every $\mathbf{c} \in \mathbb{R}^r$, there exists $\mathbf{c}' \in \mathcal{S}'$ such that $\|\mathbf{c} - \mathbf{c}'\|_2 \leq 1/2$. Then, we have

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

Taking sup and max on both side yields,

$$\sup_{\mathbf{c} \in \mathcal{S}} |\langle \mathbf{c}, \mathbf{x} \rangle| \leq 2 \max_{\mathbf{c}' \in \mathcal{S}'} |\langle \mathbf{c}', \mathbf{x} \rangle|.$$

Finally applying this result to (5) yields,

$$\begin{aligned} \mathbb{P} \left(\sup_{\mathcal{T} \in \mathcal{S}} |\langle \mathcal{T}, \mathcal{X} \rangle| \geq t \right) &\leq \sum_{\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}} \mathbb{P} \left(\max_{\mathbf{c}' \in \mathcal{S}'} \langle \mathbf{c}', \mathbf{x} \rangle \geq t/2 \right) \\ &\leq (k^{2n}) 6^r \phi(t/2) \\ &\leq \exp(Ck^m + 2n \log k) \phi(t/2), \end{aligned}$$

where the last inequality used (4) and the fact $r \leq k^m$.

□

3.2 Proof of Thoerm 2.2

Proof. First, we prove the probability tail bound. By Lemma 2.1, we can always find a block tensor Θ^* close to the true probability tensor Θ^{true} such that

$$\frac{1}{n^m} \|\Theta^* - \Theta^{\text{true}}\|_F^2 \leq m^2 L^2 \left(\frac{1}{k^2} \right)^\alpha. \quad (6)$$

By triangular inequality,

$$\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2 \leq 2 \underbrace{\|\hat{\Theta} - \Theta^*\|_F^2}_{(i)} + 2 \underbrace{\|\Theta^* - \Theta^{\text{true}}\|_F^2}_{(ii)}. \quad (7)$$

Since we have the error bound (ii) as in (6), we find the upper bound of the error (i). Based on definition of $\hat{\Theta}$, we have the following inequality similar to (3).

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

It suffices to bound the inner product term because of (6). 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

$$\mathbb{P} \left(\left\langle \frac{\hat{\Theta} - \Theta^*}{\|\hat{\Theta} - \Theta^*\|_F}, \frac{\mathcal{A} - \rho \Theta^{\text{true}}}{\rho} \right\rangle \geq t \right) \leq \exp(C' (k^m + n \log k)) \cdot \exp \left(- \min \left(\frac{\rho t^2}{24}, \frac{t \sqrt{C \rho (k^m + n \log k)}}{4} \right) \right),$$

for some universal constants $C, C' > 0$. Setting $t = \sqrt{C''(k^m + n \log k)/\rho}$ for sufficiently large C'' depending $C > 0$ yields

$$(i) \lesssim m^2 L^2 \left(\frac{1}{k} \right)^{2\alpha} + \frac{1}{\rho} \left(\left(\frac{k}{n} \right)^m + \frac{\log k}{n^{m-1}} \right)$$

with probability at least $1 - \exp(-C_2(n \log k + k^m))$. Combinations of two error bounds in (7) and setting $k = \left\lceil (\rho n^m)^{\frac{1}{m+2\alpha}} \right\rceil$, completes the theorem.

The moment bound follows by the probability tail bound. □

References

Chao Gao, Yu Lu, Zongming Ma, and Harrison H Zhou. Optimal estimation and completion of matrices with biclustering structures. *The Journal of Machine Learning Research*, 17(1):5602–5630, 2016.