A Proof of the Main Theory

In this section, we will provide a detailed proof for the main theory. We first introduce the resampling version of our proposed Algorithm 1, which is displayed in Algorithm 3.

Algorithm 3 Gradient Descent with Hard Thresholding (Resampling Version)

1: **Input:** Number of iterations T, sparsity s_1, s_2, step size η_1, η_2
2: Split the Dataset into T Subsets of Size n/T
3: **for** $t = 0$ to $T-1$ **do**
4: **Update $\boldsymbol{\Gamma}$ with the t-th Data Subset:**
$\boldsymbol{\Gamma}^{(t+0.5)} = \boldsymbol{\Gamma}^{(t)} - \eta_1 \nabla_1 f_{n/T}\big(\boldsymbol{\Gamma}^{(t)}, \boldsymbol{\Omega}^{(t)}\big)$
$\boldsymbol{\Gamma}^{(t+1)} = \text{HT}(\boldsymbol{\Gamma}^{(t+0.5)}, s_1)$
5: **Update $\boldsymbol{\Omega}$ with the t-th Data Subset:**
$\boldsymbol{\Omega}^{(t+0.5)} = \boldsymbol{\Omega}^{(t)} - \eta_2 \nabla_2 f_{n/T}\big(\boldsymbol{\Gamma}^{(t)}, \boldsymbol{\Omega}^{(t)}\big)$
$\boldsymbol{\Omega}^{(t+1)} = \text{HT}(\boldsymbol{\Omega}^{(t+0.5)}, s_2)$
6: **end for**
7: **Output:** $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^{(T)}, \widehat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}^{(T)}$

Before we begin our proof, we first define $\mathbb{B}_F(\boldsymbol{\Gamma}^*; r) = \{\boldsymbol{\Gamma} \in \mathbb{R}^{d\times m} : \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*\|_F \le r\}$. Similarly we define $\mathbb{B}_F(\boldsymbol{\Omega}^*; r) = \{\boldsymbol{\Omega} \in \mathbb{R}^{m\times m} : \|\boldsymbol{\Omega} - \boldsymbol{\Omega}^*\|_F \le r\}$. Now we introduce several lemmas, which are essential to the proof.

Lemma A.1. Under Assumptions 4.1 and 4.2, for any $\boldsymbol{\Gamma}', \boldsymbol{\Gamma} \in \mathbb{B}_F(\boldsymbol{\Gamma}^*; r)$, the population loss function $f(\cdot, \boldsymbol{\Omega}^*)$ is $2/(\nu\tau)$-strongly convex and $2\nu\tau$-smooth, i.e.,

$$\frac{1}{\nu\tau}\|\boldsymbol{\Gamma}' - \boldsymbol{\Gamma}\|_F^2 \le f(\boldsymbol{\Gamma}', \boldsymbol{\Omega}^*) - f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^*) - \langle \nabla_1 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^*), \boldsymbol{\Gamma}' - \boldsymbol{\Gamma}\rangle \le \nu\tau\|\boldsymbol{\Gamma}' - \boldsymbol{\Gamma}\|_F^2.$$

Lemma A.2. Under Assumptions 4.1 and 4.2, for any $\boldsymbol{\Omega}', \boldsymbol{\Omega} \in \mathbb{B}_F(\boldsymbol{\Omega}^*; r)$, and if $r \le 1/(2\nu)$, then the population loss function $f(\boldsymbol{\Gamma}^*, \cdot)$ is $1/(4\nu^2)$-strongly convex and $4\nu^2$-smooth, i.e.,

$$\frac{1}{8\nu^2}\|\boldsymbol{\Omega}' - \boldsymbol{\Omega}\|_F^2 \le f(\boldsymbol{\Gamma}^*, \boldsymbol{\Omega}') - f(\boldsymbol{\Gamma}^*, \boldsymbol{\Omega}) - \langle \nabla_2 f(\boldsymbol{\Gamma}^*, \boldsymbol{\Omega}), \boldsymbol{\Omega}' - \boldsymbol{\Omega}\rangle \le 2\nu^2\|\boldsymbol{\Omega}' - \boldsymbol{\Omega}\|_F^2.$$

Lemmas A.1 and A.2 indicate that when one of the two variables (i.e., $\boldsymbol{\Gamma}$ or $\boldsymbol{\Omega}$) is fixed as true variable (i.e., $\boldsymbol{\Gamma}^*$ or $\boldsymbol{\Omega}^*$), the population function f is both strongly convex and smooth with respect to the other variable. These conclusions ensure that the standard convex optimization results for strongly convex and smooth objective functions (Nesterov, 2004) can be applied to function f as long as one of the variables takes its true value.

Lemma A.3. Suppose Assumptions 4.1 and 4.2 hold. For the true parameter $\boldsymbol{\Omega}^*$ and any $\boldsymbol{\Omega} \in \mathbb{B}_F(\boldsymbol{\Omega}^*; r)$, the gradient difference $\nabla_1 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^*) - \nabla_1 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega})$ satisfies

$$\|\nabla_1 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^*) - \nabla_1 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega})\|_F \le 2\tau r \cdot \|\boldsymbol{\Omega}^* - \boldsymbol{\Omega}\|_F. \tag{A.1}$$

For true parameter $\boldsymbol{\Gamma}^*$ and any $\boldsymbol{\Gamma} \in \mathbb{B}_F(\boldsymbol{\Gamma}^*; r)$, the gradient difference $\nabla_2 f(\boldsymbol{\Gamma}^*, \boldsymbol{\Omega}) - \nabla_2 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega})$ satisfies

$$\|\nabla_2 f(\boldsymbol{\Gamma}^*, \boldsymbol{\Omega}) - \nabla_2 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega})\|_F \le \tau r \cdot \|\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma}\|_F. \tag{A.2}$$

Lemma A.3 suggests the gradients satisfy Lipschitz property with respect to $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$. Note that this Lipschitz property only holds between the true parameter ($\boldsymbol{\Gamma}^*$ or $\boldsymbol{\Omega}^*$) and arbitrary parameter in the neighborhood of the true parameter ($\boldsymbol{\Gamma} \in \mathbb{B}_F(\boldsymbol{\Gamma}^*; r)$ or $\boldsymbol{\Omega} \in \mathbb{B}_F(\boldsymbol{\Omega}^*; r)$). Given Lemma A.3, standard convex optimization results can be adapted to analyze $f(\cdot, \boldsymbol{\Omega})$ for any $\boldsymbol{\Omega} \in \mathbb{B}_F(\boldsymbol{\Omega}^*; r)$ and $f(\boldsymbol{\Gamma}, \cdot)$ for any $\boldsymbol{\Gamma} \in \mathbb{B}_F(\boldsymbol{\Gamma}^*; r)$.

The next lemma characterizes the difference between the gradients of the population and sample loss functions, in terms of $\ell_{\infty,\infty}$ norm.

Lemma A.4. For any fixed $\boldsymbol{\Gamma} \in \mathbb{B}_F(\boldsymbol{\Gamma}^*; r)$ and $\boldsymbol{\Omega} \in \mathbb{B}_F(\boldsymbol{\Omega}^*; r)$ with $r \le \min\{M, \sqrt{\nu/\tau}\}$, then with probability at least $1-\delta$ we have

$$\|\nabla_1 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}) - \nabla_1 f_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega})\|_{\infty,\infty} \le \epsilon_1(n, \delta). \tag{A.3}$$

If we choose $\delta = 2/d$ then we have $\epsilon_1(n,\delta) = CM\sqrt{\tau\nu}\sqrt{\log(dm)/n}$. Also with probability at least $1-\delta$ we have

$$\|\nabla_2 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}) - \nabla_2 f_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega})\|_{\infty,\infty} \le \epsilon_2(n, \delta). \tag{A.4}$$

If we choose $\delta = C''/m$ then we have $\epsilon_2(n,\delta) = C'M\sqrt{(\log m)/n}$.

We further define the gradient descent update for the population loss:

$$\bar{\mathbf{\Gamma}}^{(t+0.5)} = \mathbf{\Gamma}^{(t)} - \eta_1 \nabla_1 f\big(\mathbf{\Gamma}^{(t)}, \mathbf{\Omega}^{(t)}\big),\ \bar{\mathbf{\Omega}}^{(t+0.5)} = \mathbf{\Gamma}^{(t)} - \eta_2 \nabla_2 f\big(\mathbf{\Gamma}^{(t)}, \mathbf{\Omega}^{(t)}\big).$$

Our subsequent two lemmas bridge the gap between population loss update and sample loss update.

Lemma A.5. Under Assumptions 4.1 and 4.2, suppose that $\mathbf{\Gamma} \in \mathbb{B}_F(\mathbf{\Gamma}^*; r)$, then Algorithm 3 with step sizes $\eta_1 = \nu\tau/(\nu^2\tau^2+1)$ satisfies

$$\big\|\bar{\mathbf{\Gamma}}^{(t+0.5)} - \mathbf{\Gamma}^*\big\|_F \le \frac{\nu^2\tau^2-1}{\nu^2\tau^2+1} \cdot \big\|\mathbf{\Gamma}^{(t)} - \mathbf{\Gamma}^*\big\|_F + \frac{2r\nu\tau^2}{\nu^2\tau^2+1} \cdot \big\|\mathbf{\Omega}^{(t)} - \mathbf{\Omega}^*\big\|_F.$$

Similarly, we have the following lemma establishing the result for of $\|\bar{\mathbf{\Omega}}^{(t+0.5)} - \mathbf{\Omega}^*\|_F$.

Lemma A.6. Under Assumptions 4.1 and 4.2, suppose that $\mathbf{\Omega} \in \mathbb{B}_F(\mathbf{\Omega}^*; r)$, then Algorithm 3 with step size $\eta_2 = 8\nu^2/(16\nu^4+1)$ satisfies

$$\big\|\bar{\mathbf{\Omega}}^{(t+0.5)} - \mathbf{\Omega}^*\big\|_F \le \frac{16\nu^4-1}{16\nu^4+1} \cdot \big\|\mathbf{\Omega}^{(t)} - \mathbf{\Omega}^*\big\|_F + \frac{8\tau\nu^2 r}{16\nu^4+1} \cdot \big\|\mathbf{\Gamma}^{(t)} - \mathbf{\Gamma}^*\big\|_F.$$

The next lemma characterizes the effects of hard thresholding.

Lemma A.7 (Li et al. (2016))**.** Let $\boldsymbol{\beta}^*$ be a sparse vector such that $\|\boldsymbol{\beta}^*\|_0 \le s^*$, and HT be the hard thresholding operator, which keeps the largest s entries (in magnitude) and sets the other entries equal to zero. Given $s \ge s^*$, for any vector $\boldsymbol{\beta}$, we have,

$$\big\|\mathcal{HT}(\boldsymbol{\beta}, s) - \boldsymbol{\beta}^*\big\|_2^2 \le \left(1 + \frac{2\sqrt{s^*}}{\sqrt{s-s^*}}\right) \cdot \big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\|_2^2. \tag{A.5}$$

The following two lemmas demonstrate the initialization results for $\mathbf{\Gamma}^{\text{init}}$ and $\mathbf{\Omega}^{\text{init}}$.

Lemma A.8. Under Assumption 4.2, if we select the regularization parameter λ_Γ in Algorithm 3 as $\lambda_\Gamma = c_0(\tau\sqrt{ds_1^*}\log(dm)/n)^{1/3}$, then with probability at least $1 - c_1\exp(-c_2 dm)$, it holds that

$$\big\|\mathbf{\Gamma}^{\text{init}} - \mathbf{\Gamma}^*\big\|_F \le C\big(\tau(ds_1^*)^2 \cdot \log(dm)/n\big)^{1/3}. \tag{A.6}$$

Lemma A.9. Under Assumptions 4.1 and 4.2, suppose the sample size n is large enough such that $\big\|\mathbf{\Gamma}^{\text{init}} - \mathbf{\Gamma}^*\big\|_F \le \sqrt{\nu/\tau}$, if we select the regularization parameter λ_Ω in Algorithm 3 as $\lambda_\Omega = c_0 M\nu\sqrt{\log m/n} + c_1 M\nu^{\frac{1}{2}}\tau^{5/6}(ds_1^*)^{2/3} \cdot (\log dm/n)^{5/6}$, then with probability at least $1 - c_2/m$, it holds that

$$\|\mathbf{\Omega}^{\text{init}} - \mathbf{\Omega}^*\|_F \le C' M\nu\sqrt{\frac{ms_2^* \cdot \log m}{n}} + C'' M\nu^{\frac{1}{2}}\tau^{\frac{5}{6}}(ms_2^*)^{\frac{1}{2}}(ds_1^*)^{\frac{2}{3}} \cdot \left(\frac{\log dm}{n}\right)^{\frac{5}{6}}.$$

Now we have gathered everything we need and we are ready to present the proof of the main theorem.

A.1 Proof of Theorem 4.3

Proof of Theorem 4.3. We first prove that the estimation error can be controlled by R in each step, by induction. Since the initialization estimator already satisfies $\max\{\|\mathbf{\Gamma}^{(0)} - \mathbf{\Gamma}^*\|_F, \|\mathbf{\Omega}^{(0)} - \mathbf{\Omega}^*\|_F\} \le R$, We only need to prove that the estimation error in any iterate t also satisfies the above condition given the information about $(t-1)$-th iteration.

Define $\mathcal{I}_1^* = \text{supp}(\mathbf{\Gamma}^*)$, $\mathcal{I}_1^{(t)} = \text{supp}(\mathbf{\Gamma}^{(t)})$, $\mathcal{I}_1^{(t+1)} = \text{supp}(\mathbf{\Gamma}^{(t+1)})$ and $\mathcal{I}_1 = \mathcal{I}_1^* \cup \mathcal{I}_1^{(t)} \cup \mathcal{I}_1^{(t+1)}$. It is easy to verify that

$$\mathbf{\Gamma}^{(t+1)} = \mathcal{HT}(\mathbf{\Gamma}^{(t+0.5)}, s_1) = \mathcal{HT}(\mathbf{\Gamma}^{(t)} - \eta_1[\nabla_1 f_{n/T}\big(\mathbf{\Gamma}, \mathbf{\Omega}\big)]_{\mathcal{I}_1}, s_1).$$

Since we already have that $\mathbf{\Gamma}^{(0)} \in \mathbb{B}_F(\mathbf{\Gamma}^*; R)$, $\mathbf{\Omega}^{(0)} \in \mathbb{B}_F(\mathbf{\Omega}^*; R)$ by definition, now we consider expanding this using mathematical induction. Suppose that $\mathbf{\Gamma}^{(t-1)} \in \mathbb{B}_F(\mathbf{\Gamma}^*; R)$, $\mathbf{\Omega}^{(t-1)} \in \mathbb{B}_F(\mathbf{\Omega}^*; R)$. Consider the estimation error of t-th iteration, by Lemma A.7, we have

$$\begin{aligned}\big\|\mathbf{\Gamma}^{(t+1)}-\mathbf{\Gamma}^*\big\|_F&=\big\|\mathbf{\Gamma}^{(t)}-\eta_1\big[\nabla_1 f_{n/T}(\mathbf{\Gamma},\mathbf{\Omega})\big]_{\mathcal{I}_1}-\mathbf{\Gamma}^*\big\|_F\\&\le\bigg(1+\frac{2\sqrt{ds_1^*}}{\sqrt{s_1-ds_1^*}}\bigg)^{1/2}\big\|\mathbf{\Gamma}^{(t)}-\eta_1\big[\nabla_1 f_{n/T}(\mathbf{\Gamma},\mathbf{\Omega})\big]_{\mathcal{I}_1}-\mathbf{\Gamma}^*\big\|_F\\&\le\bigg(1+\frac{2\sqrt{ds_1^*}}{\sqrt{s_1-ds_1^*}}\bigg)^{1/2}\big\|\bar{\mathbf{\Gamma}}^{(t+0.5)}-\mathbf{\Gamma}^*\big\|_F+\eta_1\bigg(1+\frac{2\sqrt{ds_1^*}}{\sqrt{s_1-ds_1^*}}\bigg)^{1/2}\big\|\big[\nabla_1 f(\mathbf{\Gamma},\mathbf{\Omega})-\nabla_1 f_{n/T}(\mathbf{\Gamma},\mathbf{\Omega})\big]_{\mathcal{I}_1}\big\|_F,\end{aligned}\tag{A.7}$$

where the last inequality holds due to triangle inequality. Notice that by Lemma A.4 we have

$$\big\|\big[\nabla_1 f(\mathbf{\Gamma},\mathbf{\Omega})-\nabla_1 f_{n/T}(\mathbf{\Gamma},\mathbf{\Omega})\big]_{\mathcal{I}_1}\big\|_F\le\sqrt{|\mathcal{I}_1|}\big\|\nabla_1 f(\mathbf{\Gamma},\mathbf{\Omega})-\nabla_1 f_{n/T}(\mathbf{\Gamma},\mathbf{\Omega})\big\|_{\infty,\infty}\le\sqrt{ds_1^*+2s_1}\cdot\epsilon_1(n/T,\delta/T).$$

Therefore, (A.7) can be further written as:

$$\begin{aligned}\big\|\mathbf{\Gamma}^{(t+1)}-\mathbf{\Gamma}^*\big\|_F&\le\bigg(1+\frac{2\sqrt{ds_1^*}}{\sqrt{s_1-ds_1^*}}\bigg)^{1/2}\big\|\bar{\mathbf{\Gamma}}^{(t+0.5)}-\mathbf{\Gamma}^*\big\|_F+\eta_1\sqrt{ds_1^*+2s_1}\bigg(1+\frac{2\sqrt{ds_1^*}}{\sqrt{s_1-ds_1^*}}\bigg)^{1/2}\epsilon_1(n/T,\delta/T)\\&\le\bigg(1+\frac{2\sqrt{ds_1^*}}{\sqrt{s_1-ds_1^*}}\bigg)^{1/2}\bigg[\frac{\nu^2\tau^2-1}{\nu^2\tau^2+1}\cdot\big\|\mathbf{\Gamma}^{(t)}-\mathbf{\Gamma}^*\big\|_F+\frac{2R\nu\tau^2}{\nu^2\tau^2+1}\cdot\big\|\mathbf{\Omega}^{(t)}-\mathbf{\Omega}^*\big\|_F\bigg]\\&\quad+CM\eta_1\bigg(1+\frac{2\sqrt{ds_1^*}}{\sqrt{s_1-ds_1^*}}\bigg)^{1/2}\cdot\sqrt{\nu\tau}\sqrt{(ds_1^*+2s_1)\log(dmT)\cdot T/n},\end{aligned}\tag{A.8}$$

where the last inequality is due to Lemma A.5 and choosing $\delta=2/d$ for $\epsilon_1(n/T,\delta/T)$ in Lemma A.4. Similarly, we can also define $\mathcal{I}_2^*=\text{supp}(\mathbf{\Omega}^*)$, $\mathcal{I}_2^{(t)}=\text{supp}(\mathbf{\Omega}^{(t)})$, $\mathcal{I}_2^{(t+1)}=\text{supp}(\mathbf{\Omega}^{(t+1)})$, $\mathcal{I}_2=\mathcal{I}_2^*\cup\mathcal{I}_2^{(t)}\cup\mathcal{I}_2^{(t+1)}$, and then establish the bound for $\|\mathbf{\Omega}^{(t+1)}-\mathbf{\Omega}^*\|_F$ as:

$$\begin{aligned}\|\mathbf{\Omega}^{(t+1)}-\mathbf{\Omega}^*\|_F&\le\bigg(1+\frac{2\sqrt{ms_2^*}}{\sqrt{s_2-ms_2^*}}\bigg)^{1/2}\cdot\bigg[\frac{16\nu^4-1}{16\nu^4+1}\cdot\big\|\mathbf{\Omega}^{(t)}-\mathbf{\Omega}^*\big\|_F+\frac{8\tau\nu^2R}{16\nu^4+1}\cdot\big\|\mathbf{\Gamma}^{(t)}-\mathbf{\Gamma}^*\big\|_F\bigg]\\&\quad+C'M\eta_2\bigg(1+\frac{2\sqrt{ms_2^*}}{\sqrt{s_2-ms_2^*}}\bigg)^{1/2}\cdot\sqrt{(ms_2^*+2s_2)\log(mT)\cdot T/n}.\end{aligned}\tag{A.9}$$

Now we define

$$\begin{aligned}\rho&=\max\bigg\{\frac{\nu^2\tau^2-1}{\nu^2\tau^2+1}+\frac{2R\nu\tau^2}{\nu^2\tau^2+1},\frac{16\nu^4-1}{16\nu^4+1}+\frac{8\tau\nu^2R}{16\nu^4+1}\bigg\}\\&=\max\bigg\{1-\frac{2-2R\nu\tau^2}{\nu^2\tau^2+1},1-\frac{2-8\tau\nu^2R}{16\nu^4+1}\bigg\}.\end{aligned}$$

Note that by our assumptions $s_1\ge\big(1+4/(1/\rho-1)^2\big)ds_1^*$ and $s_2\ge\big(1+4/(1/\rho-1)^2\big)ms_2^*$, we have

$$\max\bigg\{\bigg(1+\frac{2\sqrt{ds_1^*}}{\sqrt{s_1-ds_1^*}}\bigg)^{1/2},\bigg(1+\frac{2\sqrt{ms_2^*}}{\sqrt{s_2-ms_2^*}}\bigg)^{1/2}\bigg\}\le\frac{1}{\sqrt{\rho}}.$$

Thus by combining (A.8) together with (A.9), we get

$$\max\big\{\|\mathbf{\Gamma}^{(t+1)}-\mathbf{\Gamma}^*\|_F,\|\mathbf{\Omega}^{(t+1)}-\mathbf{\Omega}^*\|_F\big\}\le\sqrt{\rho}\cdot\max\big\{\|\mathbf{\Gamma}^{(t)}-\mathbf{\Gamma}^*\|_F,\|\mathbf{\Omega}^{(t)}-\mathbf{\Omega}^*\|_F\big\}+\max\big\{\alpha_1,\alpha_2\big\},\tag{A.10}$$

where α_1 and α_2 are defined as

$$\alpha_1=C'M\sqrt{\nu\tau}\cdot\sqrt{\frac{ds_1^*\log(dmT)}{n/T}},\qquad\alpha_2=C''M\cdot\sqrt{\frac{ms_2^*\log(mT)}{n/T}}.\tag{A.11}$$

For simplicity, if we denote $P^{(t)} = \max\big\{\|\mathbf{\Gamma}^{(t)} - \mathbf{\Gamma}^*\|_F, \|\mathbf{\Omega}^{(t)} - \mathbf{\Omega}^*\|_F\big\}$, $\zeta = \max\{\alpha_1, \alpha_2\}$ and take one step back from iteration $t+1$ to t, then (A.10) can be rewritten as:

$$P^{(t)} \le \sqrt{\rho} \cdot P^{(t-1)} + \zeta. \tag{A.12}$$

Since we have $\mathbf{\Gamma}^{(t-1)} \in \mathbb{B}_F(\mathbf{\Gamma}^*; R)$, $\mathbf{\Omega}^{(t-1)} \in \mathbb{B}_F(\mathbf{\Omega}^*; R)$, by (A.12), it immediately implies that

$$P^{(t-1)} = \max\big\{\|\mathbf{\Gamma}^{(t-1)} - \mathbf{\Gamma}^*\|_F, \|\mathbf{\Omega}^{(t-1)} - \mathbf{\Omega}^*\|_F\big\} \le R.$$

Given the theorem condition (4.1), we can easily derive that

$$\zeta \le \big(1 - \sqrt{\rho}\big)R.$$

Thus we have

$$P^{(t)} \le \sqrt{\rho} \cdot P^{(t-1)} + \zeta \le \sqrt{\rho} \cdot R + r(1 - \sqrt{\rho}) \le R.$$

Therefore we proved that for all $t \ge 1$, $\mathbf{\Gamma}^{(t)} \in \mathbb{B}_F(\mathbf{\Gamma}^*; R)$, $\mathbf{\Omega}^{(t)} \in \mathbb{B}_F(\mathbf{\Omega}^*; R)$.
Next we prove the bound in the theorem. Consider

$$\begin{aligned} P^{(t)} &\le \sqrt{\rho} \cdot P^{(t-1)} + \zeta \le \rho \cdot P^{(t-2)} + \sqrt{\rho} \cdot \zeta + \zeta \le ... \\ &\le \rho^{t/2} \cdot P^{(0)} + \rho^{(t-1)/2} \cdot \zeta + ... + \zeta \\ &\le \rho^{t/2} \cdot P^{(0)} + \frac{1}{1 - \sqrt{\rho}}\zeta, \end{aligned}$$

where the last inequality holds for series summation rule when $t \to \infty$. Since $P^{(0)} = r \le R$, we rewrite the above inequality as

$$\max\big\{\|\mathbf{\Gamma}^{(t)} - \mathbf{\Gamma}^*\|_F, \|\mathbf{\Omega}^{(t)} - \mathbf{\Omega}^*\|_F\big\} \le \rho^{t/2} \cdot R + \frac{1}{1 - \sqrt{\rho}}\zeta, \text{ for all } t \in [T].$$

This completes the proof. □

A.2 Proof of Theorem 4.7

In this section, we present the analysis of our initialization algorithm (Algorithm 2). The main idea in this analysis is inspired from Yang et al. (2014a;b) for elementary Gaussian graphical models. However, in our initialization estimator, we use ridge type graphical model estimator rather than performing diagonal enhancement operator on the sample covariance matrix as in Yang et al. (2014b). Therefore, for the self-containedness of our paper, we choose to present the proof here.

Proof of Theorem 4.7. According to Lemma A.8, we have that

$$\|\mathbf{\Gamma}^{\text{init}} - \mathbf{\Gamma}^*\|_F \le C_m\big(\tau(ds_1^*)^2 \cdot \log(dm)/n\big)^{1/3}$$

Thus according to the theorem condition on the sample size n, we can easily get

$$\|\mathbf{\Gamma}^{\text{init}} - \mathbf{\Gamma}^*\|_F \le R/2.$$

The same argument applies to the initial estimator $\mathbf{\Omega}^{\text{init}}$. According to Lemma A.9, we have that

$$\|\mathbf{\Omega}^{\text{init}} - \mathbf{\Omega}^*\|_F \le C_{g1}M\nu\sqrt{\frac{ms_2^* \cdot \log m}{n}} + C_{g2}M\nu^{\frac{1}{2}}\tau^{\frac{5}{6}}(ms_2^*)^{\frac{1}{2}}(ds_1^*)^{\frac{2}{3}} \cdot \left(\frac{\log dm}{n}\right)^{\frac{5}{6}}.$$

Therefore, according to theorem condition on the sample size n, we can easily have

$$\|\mathbf{\Omega}^{\text{init}} - \mathbf{\Omega}^*\|_F \le R/2.$$

By Lemma A.7, for any $s_1 \geq 4ds_1^*$ we have

$$\|\mathbf{\Gamma}^{(0)} - \mathbf{\Gamma}^*\|_F \leq \bigg(1 + \frac{2\sqrt{ds_1^*}}{\sqrt{s_1 - ds_1^*}}\bigg)^{1/2} \cdot \|\mathbf{\Gamma}^{\text{init}} - \mathbf{\Gamma}^*\|_F \leq \bigg(1 + \frac{2\sqrt{ds_1^*}}{\sqrt{s_1 - ds_1^*}}\bigg)^{1/2} \frac{R}{2} \leq R.$$

Similarly, we can prove that for $\mathbf{\Omega}$, we have $\|\mathbf{\Omega}^{(0)} - \mathbf{\Omega}^*\|_F \leq R$. Thus we prove that by initialization, the initial estimation error satisfies

$$\max\{\|\mathbf{\Gamma}^{(0)} - \mathbf{\Gamma}^*\|_F, \|\mathbf{\Omega}^{(0)} - \mathbf{\Omega}^*\|_F\} \leq R.$$

□

B Proof of Technical Lemmas in Section A

In the following, we will give detailed proof of the technical lemmas used in Section A. First let us denote $\boldsymbol{\epsilon}_i = \mathbf{y}_i - \mathbf{\Gamma}^{*\top}\mathbf{x}_i$ for the rest of this section. Note that from the CGGM model we can obtain that $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{\Omega}^{*-1})$.

B.1 Proof of Lemma A.1

Proof. Recall that

$$\begin{aligned} f_n(\mathbf{\Gamma}, \mathbf{\Omega}) &= -\log|\mathbf{\Omega}| + \frac{1}{n}\sum_{i=1}^n \big((\mathbf{y}_i - \mathbf{\Gamma}^\top\mathbf{x}_i)^\top \mathbf{\Omega}(\mathbf{y}_i - \mathbf{\Gamma}^\top\mathbf{x}_i)\big) \\ &= -\log|\mathbf{\Omega}| + \frac{1}{n}\sum_{i=1}^n \big((\mathbf{\Gamma}^{*\top}\mathbf{x}_i - \mathbf{\Gamma}^\top\mathbf{x}_i + \boldsymbol{\epsilon}_i)^\top \mathbf{\Omega}(\mathbf{\Gamma}^{*\top}\mathbf{x}_i - \mathbf{\Gamma}^\top\mathbf{x}_i + \boldsymbol{\epsilon}_i)\big). \end{aligned} \tag{B.1}$$

Based on the above equality we compute the population version of f function:

$$\begin{aligned} f(\mathbf{\Gamma}, \mathbf{\Omega}) &= \mathbb{E}\big[f_n(\mathbf{\Gamma}, \mathbf{\Omega})\big] \\ &= \mathbb{E}\bigg[-\log|\mathbf{\Omega}| + \frac{1}{n}\sum_{i=1}^n \big((\mathbf{\Gamma}^{*\top}\mathbf{x}_i - \mathbf{\Gamma}^\top\mathbf{x}_i + \boldsymbol{\epsilon}_i)^\top \mathbf{\Omega}(\mathbf{\Gamma}^{*\top}\mathbf{x}_i - \mathbf{\Gamma}^\top\mathbf{x}_i + \boldsymbol{\epsilon}_i)\big)\bigg] \\ &= -\log|\mathbf{\Omega}| + \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^\top(\mathbf{\Gamma}^* - \mathbf{\Gamma})\mathbf{\Omega}(\mathbf{\Gamma}^* - \mathbf{\Gamma})^\top\mathbf{x}_i\bigg] + \text{tr}(\mathbf{\Omega}\mathbf{\Omega}^{*-1}) \\ &= -\log|\mathbf{\Omega}| + \text{tr}\big((\mathbf{\Gamma}^* - \mathbf{\Gamma})\mathbf{\Omega}(\mathbf{\Gamma}^* - \mathbf{\Gamma})^\top\mathbf{\Sigma}_X^*\big) + \text{tr}(\mathbf{\Omega}\mathbf{\Omega}^{*-1}). \end{aligned} \tag{B.2}$$

Thus, we get

$$\nabla_1 f(\mathbf{\Gamma}, \mathbf{\Omega}) = -2\mathbf{\Sigma}_X^*(\mathbf{\Gamma}^* - \mathbf{\Gamma})\mathbf{\Omega}. \tag{B.3}$$

Apply vectorization and use the property of Kronecker product that $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\text{vec}(\mathbf{B})$, we obtain

$$\nabla_1^2 f(\mathbf{\Gamma}, \mathbf{\Omega}^*) = \mathbf{\Omega}^* \otimes 2\mathbf{\Sigma}_X^*.$$

For function $f(\cdot, \mathbf{\Omega}^*)$, according to Taylor expansion, we have

$$\begin{aligned} f(\mathbf{\Gamma}', \mathbf{\Omega}^*) &= f(\mathbf{\Gamma}, \mathbf{\Omega}^*) + \langle \nabla_1 f(\mathbf{\Gamma}, \mathbf{\Omega}^*), \mathbf{\Gamma}' - \mathbf{\Gamma}\rangle \\ &\quad + \frac{1}{2}\langle \text{vec}(\mathbf{\Gamma}') - \text{vec}(\mathbf{\Gamma}), \nabla_1^2 f(\mathbf{\Gamma}, \mathbf{\Omega}^*)(\text{vec}(\mathbf{\Gamma}') - \text{vec}(\mathbf{\Gamma}))\rangle. \end{aligned} \tag{B.4}$$

Then (B.4) further implies

$$\begin{aligned} f(\mathbf{\Gamma}', \mathbf{\Omega}^*) - f(\mathbf{\Gamma}, \mathbf{\Omega}^*) - \langle \nabla_1 f(\mathbf{\Gamma}, \mathbf{\Omega}^*), \mathbf{\Gamma}' - \mathbf{\Gamma}\rangle &\leq \frac{1}{2}\lambda_{\max}(\nabla_1^2 f(\mathbf{\Gamma}, \mathbf{\Omega}^*))\|\mathbf{\Gamma}' - \mathbf{\Gamma}\|_F^2, \\ f(\mathbf{\Gamma}', \mathbf{\Omega}^*) - f(\mathbf{\Gamma}, \mathbf{\Omega}^*) - \langle \nabla_1 f(\mathbf{\Gamma}, \mathbf{\Omega}^*), \mathbf{\Gamma}' - \mathbf{\Gamma}\rangle &\geq \frac{1}{2}\lambda_{\min}(\nabla_1^2 f(\mathbf{\Gamma}, \mathbf{\Omega}^*))\|\mathbf{\Gamma}' - \mathbf{\Gamma}\|_F^2. \end{aligned}$$

Recall that $\nabla_1^2 f(\mathbf{\Gamma}, \mathbf{\Omega}^*) = 2\mathbf{\Omega}^* \otimes \widehat{\mathbf{\Sigma}}_X$ and $\|\mathbf{\Omega}^*\|_2 \leq \nu$, $\|\mathbf{\Sigma}_X^*\|_2 \leq \tau$ by Assumptions 4.1 and 4.2, we have

$$\lambda_{\max}(\nabla_1^2 f(\mathbf{\Gamma}, \mathbf{\Omega}^*)) \leq 2\nu\tau.$$

Similarly, we have

$$\lambda_{\min}(\nabla_1^2 f(\mathbf{\Gamma}, \mathbf{\Omega}^*)) \geq \frac{2}{\nu\tau}.$$

Therefore, function $f(\cdot, \mathbf{\Omega}^*)$ is $2/(\nu\tau)$-strongly convex and $2\nu\tau$-smooth function:

$$\frac{1}{\nu\tau}\|\mathbf{\Gamma}' - \mathbf{\Gamma}\|_F^2 \leq f(\mathbf{\Gamma}', \mathbf{\Omega}^*) - f(\mathbf{\Gamma}, \mathbf{\Omega}^*) - \nabla_1 f(\mathbf{\Gamma}, \mathbf{\Omega}^*)^\top(\mathbf{\Gamma}' - \mathbf{\Gamma}) \leq \nu\tau \cdot \|\mathbf{\Gamma}' - \mathbf{\Gamma}\|_F^2.$$

This completes the proof.

□

B.2 Proof of Lemma A.2

Proof. From (B.2), we have

$$\nabla_2^2 f(\mathbf{\Gamma}^*, \mathbf{\Omega}) = \mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1}.$$

According to Mean Value Theorem, we have

$$\begin{aligned} f(\mathbf{\Gamma}^*, \mathbf{\Omega}') &= f(\mathbf{\Gamma}^*, \mathbf{\Omega}) + \langle \nabla_2 f(\mathbf{\Gamma}^*, \mathbf{\Omega}), \mathbf{\Omega}' - \mathbf{\Omega}\rangle \\ &\quad + \frac{1}{2}\langle \mathrm{vec}(\mathbf{\Omega}') - \mathrm{vec}(\mathbf{\Omega}), \nabla_2^2 f(\mathbf{\Gamma}^*, \mathbf{Z})(\mathrm{vec}(\mathbf{\Omega}') - \mathrm{vec}(\mathbf{\Omega}))\rangle, \end{aligned}$$

where $\mathbf{Z} = t\mathbf{\Omega}' + (1-t)\mathbf{\Omega}$ with $t \in [0, 1]$. Define $\mathbf{\Delta} = \mathbf{\Omega}' - \mathbf{\Omega}$, we have

$$\begin{aligned} \lambda_{\min}(\nabla_2^2 f(\mathbf{\Gamma}^*, \mathbf{Z})) &= \lambda_{\min}(\mathbf{Z}^{-1} \otimes \mathbf{Z}^{-1}) = \lambda_{\min}(\mathbf{Z}^{-1})^2 = \|\mathbf{\Omega} + t\mathbf{\Delta}\|_2^{-2} \\ &\geq [\|\mathbf{\Omega}^*\|_2 + (1-t)\|\mathbf{\Omega} - \mathbf{\Omega}^*\|_2 + t\|\mathbf{\Omega}' - \mathbf{\Omega}^*\|_2]^{-2} \\ &\geq (\nu + r)^{-2} \\ &\geq \frac{1}{4\nu^2}, \end{aligned}$$

where the first inequality holds due to triangle inequality, the second inequality holds for $\|\mathbf{\Omega} - \mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega} - \mathbf{\Omega}^*\|_F \leq r$ and $\|\mathbf{\Omega}' - \mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega}' - \mathbf{\Omega}^*\|_F \leq r$ and the last inequality follows from condition $r \leq 1/(2\nu) \leq \nu/2$. Similarly, we have

$$\lambda_{\max}(\nabla_2^2 f(\mathbf{\Gamma}^*, \mathbf{Z})) = \lambda_{\max}(\mathbf{Z}^{-1} \otimes \mathbf{Z}^{-1}) = \lambda_{\max}(\mathbf{Z}^{-1})^2 = \lambda_{\min}(\mathbf{Z})^{-2}.$$

Note that $\mathbf{Z} = t\mathbf{\Omega}' + (1-t)\mathbf{\Omega} = \mathbf{\Omega} + t\mathbf{\Delta}$, and for any $\|\mathbf{x}\|_2 = 1$ we have

$$\begin{aligned} \mathbf{x}^\top(\mathbf{\Omega} + t\mathbf{\Delta})\mathbf{x} &= \mathbf{x}^\top\big(\mathbf{\Omega}^* + (1-t)(\mathbf{\Omega} - \mathbf{\Omega}^*) + t(\mathbf{\Omega}' - \mathbf{\Omega}^*)\big)\mathbf{x} \\ &\geq \mathbf{x}^\top\mathbf{\Omega}^*\mathbf{x} - (1-t)|\mathbf{x}^\top(\mathbf{\Omega} - \mathbf{\Omega}^*)\mathbf{x}| - t|\mathbf{x}^\top(\mathbf{\Omega}' - \mathbf{\Omega}^*)\mathbf{x}|, \end{aligned}$$

where the inequality holds since $t \in [0, 1]$. Taking minimization on both sides we have

$$\begin{aligned} \lambda_{\min}(\mathbf{Z}) &\geq \lambda_{\min}(\mathbf{\Omega}^*) - (1-t)|\lambda_{\max}(\mathbf{\Omega} - \mathbf{\Omega}^*)| - t|\lambda_{\max}(\mathbf{\Omega}' - \mathbf{\Omega}^*)| \\ &= \lambda_{\min}(\mathbf{\Omega}^*) - (1-t)\|\mathbf{\Omega} - \mathbf{\Omega}^*\|_2 - t\|\mathbf{\Omega}' - \mathbf{\Omega}^*\|_2 \\ &\geq \frac{1}{\nu} - r, \end{aligned}$$

where the second inequality follows from $\|\mathbf{\Omega} - \mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega} - \mathbf{\Omega}^*\|_F \leq r$ and $\|\mathbf{\Omega}' - \mathbf{\Omega}^*\|_2 \leq \|\mathbf{\Omega}' - \mathbf{\Omega}^*\|_F \leq r$. Since $r \leq 1/(2\nu)$, therefore,

$$\lambda_{\max}(\nabla_2^2 f(\mathbf{\Gamma}^*, \mathbf{Z})) \leq 4\nu^2.$$

Combining the above results, we have

$$\frac{1}{8\nu^2}\|\mathbf{\Omega}' - \mathbf{\Omega}\|_F^2 \leq f(\mathbf{\Gamma}^*, \mathbf{\Omega}') - f(\mathbf{\Gamma}^*, \mathbf{\Omega}) - \nabla_2 f(\mathbf{\Gamma}^*, \mathbf{\Omega})^\top(\mathbf{\Omega}' - \mathbf{\Omega}) \leq 2\nu^2\|\mathbf{\Omega}' - \mathbf{\Omega}\|_F^2.$$

This completes the proof. □

B.3 Proof of Lemma A.3

Proof. First, we bound $\big\|\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}^*)-\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega})\big\|_F$ in (A.1). From (B.3) we have

$$\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}) = -2\boldsymbol{\Sigma}_X^*(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})\boldsymbol{\Omega}, \quad \nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}^*) = -2\boldsymbol{\Sigma}_X^*(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})\boldsymbol{\Omega}^*.$$

Thus we get

$$\big\|\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}^*)-\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega})\big\|_F = \big\|2\boldsymbol{\Sigma}_X^*(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})(\boldsymbol{\Omega}^*-\boldsymbol{\Omega})\big\|_F \le 2\big\|\boldsymbol{\Sigma}_X^*\big\|_2\cdot\|\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma}\|_F\cdot\|\boldsymbol{\Omega}^*-\boldsymbol{\Omega}\|_F. \tag{B.5}$$

Note that we have $\lambda_{\max}\big(\boldsymbol{\Sigma}_X^*\big)\le\tau$ by Assumption 4.2, and recall the fact that $\|\boldsymbol{\Gamma}-\boldsymbol{\Gamma}^*\|_F\le r$. Therefore, (B.5) can be further bounded as

$$\big\|\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}^*)-\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega})\big\|_F \le 2\tau r\cdot\|\boldsymbol{\Omega}^*-\boldsymbol{\Omega}\|_F.$$

Now consider $\big\|\nabla_2 f(\boldsymbol{\Gamma}^*,\boldsymbol{\Omega})-\nabla_2 f(\boldsymbol{\Gamma},\boldsymbol{\Omega})\big\|_F$ in (A.2). From (B.2), we have

$$\nabla_2 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}) = -\boldsymbol{\Omega}^{-1}+\boldsymbol{\Omega}^{*-1}+(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})^\top\boldsymbol{\Sigma}_X^*(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma}), \quad \nabla_2 f(\boldsymbol{\Gamma}^*,\boldsymbol{\Omega}) = -\boldsymbol{\Omega}^{-1}+\boldsymbol{\Omega}^{*-1}.$$

Thus, we obtain

$$\big\|\nabla_2 f(\boldsymbol{\Gamma}^*,\boldsymbol{\Omega})-\nabla_2 f(\boldsymbol{\Gamma},\boldsymbol{\Omega})\big\|_F = \Big\|(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})^\top\boldsymbol{\Sigma}_X^*(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})\Big\|_F \le \|\boldsymbol{\Sigma}_X^*\|_2\cdot\|\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma}\|_F^2,$$

where the inequality follows from Cauchy-Schwartz inequality. Following similar proof procedure in (B.5), we can further bound the above inequality as

$$\big\|\nabla_2 f(\boldsymbol{\Gamma}^*,\boldsymbol{\Omega})-\nabla_2 f(\boldsymbol{\Gamma},\boldsymbol{\Omega})\big\|_F \le \tau\|\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma}\|_F^2 \le \tau r\cdot\|\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma}\|_F.$$

This completes the proof. □

B.4 Proof of Lemma A.4

Proof. **Part I:** Proof of the bound in (A.3).

Since $\boldsymbol{\epsilon}_i\sim N(\mathbf{0},\boldsymbol{\Omega}^{*-1})$, we have $\max_{ij}\|\epsilon_{ij}\|_{\psi_2}^2\le C_1\lambda_{\max}(\boldsymbol{\Omega}^{*-1})\le C_1\nu$. From (B.1), we get

$$\nabla_1 f_n(\boldsymbol{\Gamma},\boldsymbol{\Omega}) = -\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i(\mathbf{y}_i^\top-\mathbf{x}_i^\top\boldsymbol{\Gamma})\boldsymbol{\Omega} = -\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})\boldsymbol{\Omega} - \frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top\boldsymbol{\Omega}. \tag{B.6}$$

From (B.2) we have

$$\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}) = -2\boldsymbol{\Sigma}_X^*(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})\boldsymbol{\Omega}. \tag{B.7}$$

Combining (B.6) and (B.7) we obtain

$$\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega})-\nabla_1 f_n(\boldsymbol{\Gamma},\boldsymbol{\Omega}) = \frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top\boldsymbol{\Omega} + 2\bigg(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top-\boldsymbol{\Sigma}_X^*\bigg)\cdot(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})\cdot\boldsymbol{\Omega}. \tag{B.8}$$

Then we have

$$\begin{aligned}\|\nabla_1 f(\boldsymbol{\Gamma},\boldsymbol{\Omega})-\nabla_1 f_n(\boldsymbol{\Gamma},\boldsymbol{\Omega})\|_{\infty,\infty} \le\ & \underbrace{\bigg\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top(\boldsymbol{\Omega}-\boldsymbol{\Omega}^*)\bigg\|_{\infty,\infty}}_{I_1} + \underbrace{\bigg\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top\boldsymbol{\Omega}^*\bigg\|_{\infty,\infty}}_{I_2} \\ &+ \underbrace{\bigg\|2\bigg(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top-\boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})(\boldsymbol{\Omega}-\boldsymbol{\Omega}^*)\bigg\|_{\infty,\infty}}_{I_3} \\ &+ \underbrace{\bigg\|2\bigg(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top-\boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^*-\boldsymbol{\Gamma})\boldsymbol{\Omega}^*\bigg\|_{\infty,\infty}}_{I_4}.\end{aligned} \tag{B.9}$$

In the following, let $\mathbf{A}^{(i)} = \mathbf{x}_i\boldsymbol{\epsilon}_i^\top$, $\mathbf{B}^{(i)} = \big(\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\big)\cdot(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\cdot(\boldsymbol{\Omega} - \boldsymbol{\Omega}^*)$, $\mathbf{F} = \boldsymbol{\Omega} - \boldsymbol{\Omega}^*$. For term I_1, consider the ψ_1 norm of each element in $\mathbf{A}^{(i)}\mathbf{F}$:

$$\|[A^{(i)}F]_{jk}\|_{\psi_1} = \bigg\|\sum_{\ell=1}^m A_{j\ell}F_{\ell k}\bigg\|_{\psi_1} = \bigg\|\sum_{\ell=1}^m x_{ij}\epsilon_{i\ell}F_{\ell k}\bigg\|_{\psi_1}. \tag{B.10}$$

From Assumption 4.1 and Assumption 4.2 we have $\max_{ij}\|\epsilon_{ij}\|_{\psi_2}^2 \le C_1\nu$, $\max_{ij}\|x_{ij}\|_{\psi_2}^2 \le C_2\tau$, (B.10) can be further bounded as:

$$\begin{aligned}\|[A^{(i)}F]_{jk}\|_{\psi_1} &= \bigg\|x_{ij}\sum_{\ell=1}^m \epsilon_{i\ell}F_{\ell k}\bigg\|_{\psi_1} \le 2\|x_{ij}\|_{\psi_2}\cdot\bigg\|\sum_{\ell=1}^m \epsilon_{i\ell}F_{\ell k}\bigg\|_{\psi_2} \\ &\le 2\sqrt{C_1C_2\nu\tau}\sqrt{C_3\sum_{\ell=1}^m F_{\ell k}^2} \le 2\sqrt{\nu\tau}\sqrt{C_1C_2C_3}\big\|\mathbf{F}\big\|_F, \end{aligned} \tag{B.11}$$

where the first inequality follows from Lemma D.2 and the second inequality follows from Lemma D.1. Note that $\big\|\mathbf{F}\big\|_F = \|\boldsymbol{\Omega} - \boldsymbol{\Omega}^*\|_F \le r$, (B.11) can be further bounded by

$$\|[A^{(i)}F]_{jk}\|_{\psi_1} \le 2\sqrt{\nu\tau}\sqrt{C_1C_2C_3}r.$$

By Bernstein-type inequality in Theorem D.4, we have

$$\mathbb{P}\bigg(\bigg|\frac{2}{n}\sum_{i=1}^n [A^{(i)}F]_{jk}\bigg| > t\bigg) \le \exp\Big(-\frac{C_4nt^2}{16C_1C_2C_3\nu\tau r^2}\Big).$$

Applying union bound to all possible pairs of $j\in[d]$, $k\in[m]$, we get

$$\mathbb{P}\bigg\{\bigg\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top(\boldsymbol{\Omega} - \boldsymbol{\Omega}^*)\bigg\|_{\infty,\infty} \ge t\bigg\} \le \sum_{j,k}\mathbb{P}\bigg\{\bigg|\frac{2}{n}\sum_{i=1}^n [A^{(i)}F]_{jk}\bigg| \ge t\bigg\} \le d\cdot m\cdot\exp\Big(-\frac{C_4nt^2}{16C_1C_2C_3\nu\tau r^2}\Big).$$

Choose $t = 2\sqrt{\nu\tau}r\sqrt{C_1C_2C_3/C_4}\sqrt{(2\log d + \log m)/n}$, $C = 2\sqrt{C_1C_2C_3/C_4}$ and with probability at least $1 - 1/d$ we have

$$\bigg\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top(\boldsymbol{\Omega} - \boldsymbol{\Omega}^*)\bigg\|_{\infty,\infty} \le C\sqrt{\nu\tau}r\sqrt{\frac{2\log d + \log m}{n}}. \tag{B.12}$$

For term I_2, first notice that

$$\bigg\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top\boldsymbol{\Omega}^*\bigg\|_{\infty,\infty} \le \bigg\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top\bigg\|_{\infty,\infty}\cdot\|\boldsymbol{\Omega}^*\|_\infty \le M\cdot\bigg\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top\bigg\|_{\infty,\infty},$$

where the last inequality holds due to Assumption 4.1. Now consider the ψ_1 norm of each element in $\mathbf{A}^{(i)}$:

$$\|\mathbf{A}_{jk}^{(i)}\|_{\psi_1} = \|x_{ij}\epsilon_{ik}\|_{\psi_1} \le 2\|x_{ij}\|_{\psi_2}\cdot\|\epsilon_{ik}\|_{\psi_2} \le 2\sqrt{C_1C_2\nu\tau}, \tag{B.13}$$

where the first inequality follows from Lemma D.2 and the last inequality follows from Assumption 4.1 and Assumption 4.2. By Bernstein-type inequality in Theorem D.4, we have

$$\mathbb{P}\bigg(\bigg|\frac{2}{n}\sum_{i=1}^n A_{jk}^{(i)}\bigg| > t\bigg) \le \exp\Big(-\frac{C_4nt^2}{4C_1C_2\nu\tau}\Big).$$

Applying union bound to all possible pairs of $j\in[d]$, $k\in[m]$, we get

$$\mathbb{P}\bigg\{\bigg\|\frac{2}{n}\sum_{i=1}^n \mathbf{x}_i\boldsymbol{\epsilon}_i^\top\boldsymbol{\Omega}^*\bigg\|_{\infty,\infty} \ge t\bigg\} \le M\cdot\sum_{j,k}\mathbb{P}\bigg\{\bigg|\frac{2}{n}\sum_{i=1}^n A_{jk}^{(i)}\bigg| \ge t\bigg\} \le M\cdot d\cdot m\cdot\exp\Big(-\frac{C_4nt^2}{4C_1C_2\nu\tau}\Big).$$

Choose $t = 2\sqrt{\nu\tau}\sqrt{C_1C_2/C_4}\sqrt{(2\log d + \log m)/n}$, $C = 2\sqrt{C_1C_2/C_4}$ and with probability at least $1 - 1/d$ we have

$$\bigg\|\frac{2}{n}\sum_{i=1}^{n}\mathbf{x}_i\boldsymbol{\epsilon}_i^\top\boldsymbol{\Omega}^*\bigg\|_{\infty,\infty} \le C'M\cdot\sqrt{\nu\tau}\sqrt{\frac{2\log d + \log m}{n}}. \tag{B.14}$$

For term I_3, denote $(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})(\boldsymbol{\Omega} - \boldsymbol{\Omega}^*)$ as $\mathbf{G}$, $\big(\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\big)$ as $\boldsymbol{\Delta}$, note that

$$\|B_{jk}^{(i)}\|_{\psi_1} = \bigg\|\sum_{\ell=1}^{d}\Delta_{j\ell}G_{\ell k}\bigg\|_{\psi_1} = \bigg\|\sum_{\ell=1}^{d}\big[(x_{ij}x_{i\ell} - (\Sigma_X^*)_{j\ell}\big]G_{\ell k}\bigg\|_{\psi_1}. \tag{B.15}$$

Since from Assumption 4.2 we have $\max_{ij}\|x_{ij}\|_{\psi_2} \le C_5\sqrt{\tau}$, (B.15) can be further bounded as:

$$\begin{aligned}\|B_{jk}^{(i)}\|_{\psi_1} &\le 2\bigg\|x_{ij}\sum_{\ell=1}^{d}x_{i\ell}G_{\ell k}\bigg\|_{\psi_1} \le 4\big\|x_{ij}\big\|_{\psi_2}\cdot\bigg\|\sum_{\ell=1}^{d}x_{i\ell}G_{\ell k}\bigg\|_{\psi_2}\\ &\le 4C_5^2\tau\sqrt{C_6\sum_{\ell=1}^{d}G_{\ell k}^2} \le 4C_5^2\tau\sqrt{C_6}\big\|\mathbf{G}\big\|_F,\end{aligned} \tag{B.16}$$

where the first inequality follows from Lemma D.3, the second inequality holds due to Lemma D.2 and the third inequality follows from Lemma D.1. Note that $\big\|\mathbf{G}\big\|_F \le \big\|\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma}\big\|_F\cdot\big\|\boldsymbol{\Omega} - \boldsymbol{\Omega}^*\big\|_F \le r^2$, (B.16) can be further bounded by

$$\|B_{jk}^{(i)}\|_{\psi_1} \le 4C_5^2\tau\sqrt{C_6}r^2.$$

By Bernstein-type inequality in Theorem D.4, we have

$$\mathbb{P}\bigg(\bigg|\frac{2}{n}\sum_{i=1}^{n}B_{jk}^{(i)}\bigg| > t\bigg) \le \exp\Big(-\frac{C_7nt^2}{64C_5^4\tau^2C_6r^4}\Big).$$

Applying union bound to all possible pairs of $j \in [d]$, $k \in [m]$, we get

$$\mathbb{P}\bigg(\bigg\|2\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})(\boldsymbol{\Omega} - \boldsymbol{\Omega}^*)\bigg\|_{\infty,\infty} > t\bigg) \le dm\cdot\exp\Big(-\frac{C_7nt^2}{64C_5^4\tau^2C_6r^4}\Big).$$

Choose $t = 8\tau r^2\sqrt{C_5^2C_6/C_7}\sqrt{(2\log d + \log m)/n}$, $C_8 = 8\sqrt{C_5^2C_6/C_7}$, with probability at least $1 - 1/d$ we have

$$\bigg\|2\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})(\boldsymbol{\Omega} - \boldsymbol{\Omega}^*)\bigg\|_{\infty,\infty} \le C_8\tau r^2\sqrt{\frac{2\log d + \log m}{n}}. \tag{B.17}$$

For term I_4, we have

$$\begin{aligned}\bigg\|2\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\boldsymbol{\Omega}^*\bigg\|_{\infty,\infty} &\le \bigg\|2\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\bigg\|_{\infty,\infty}\cdot\big\|\boldsymbol{\Omega}^*\big\|_\infty\\ &\le M\cdot\bigg\|2\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\bigg\|_{\infty,\infty}.\end{aligned}$$

Using similar technique as we have for term I_3, with probability at least $1 - 1/d$ we have that

$$\bigg\|2\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\boldsymbol{\Omega}^*\bigg\|_{\infty,\infty} \le C_8M\tau r\sqrt{\frac{2\log d + \log m}{n}}. \tag{B.18}$$

Note that from lemma conditions we have $r \le \min\{M, \sqrt{\nu/\tau}\}$, submit (B.12), (B.17), (B.14) and (B.18) into (B.9) and apply union bound, we have with probability at least $1 - 2/d$ that

$$\|\nabla_1 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}) - \nabla_1 f_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega})\|_{\infty,\infty} \le C'M\sqrt{\nu\tau}\sqrt{\frac{2\log d + \log m}{n}}.$$

Thus we have the conclusion in (A.3).
Part II: Proof of the bound in (A.4).
We have

$$\nabla_2 f_n(\boldsymbol{\Gamma},\boldsymbol{\Omega}) = -\boldsymbol{\Omega}^{-1} + \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\Gamma}^{*\top}\mathbf{x}_i - \boldsymbol{\Gamma}^\top\mathbf{x}_i + \boldsymbol{\epsilon}_i)(\boldsymbol{\Gamma}^{*\top}\mathbf{x}_i - \boldsymbol{\Gamma}^\top\mathbf{x}_i + \boldsymbol{\epsilon}_i)^\top. \tag{B.19}$$

From (B.2) we obtain

$$\nabla_2 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}) = -\boldsymbol{\Omega}^{-1} + (\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top\boldsymbol{\Sigma}_X^*(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma}) + \boldsymbol{\Omega}^{*-1}. \tag{B.20}$$

Thus we get

$$\begin{aligned}\nabla_2 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}) - \nabla_2 f_n(\boldsymbol{\Gamma},\boldsymbol{\Omega}) = &-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top + \boldsymbol{\Omega}^{*-1} - \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top\mathbf{x}_i\boldsymbol{\epsilon}_i^\top \\ &- \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i\mathbf{x}_i^\top(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma}) + (\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma}).\end{aligned}$$

Then we have

$$\begin{aligned}\|\nabla_2 f(\boldsymbol{\Gamma},\boldsymbol{\Omega}) - \nabla_2 f_n(\boldsymbol{\Gamma},\boldsymbol{\Omega})\|_{\infty,\infty} \le &\underbrace{\bigg\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top - \boldsymbol{\Sigma}^*\bigg\|_{\infty,\infty}}_{I_1} + \underbrace{\bigg\|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top\mathbf{x}_i\boldsymbol{\epsilon}_i^\top\bigg\|_{\infty,\infty}}_{I_2} \\ &+ \underbrace{\bigg\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i\mathbf{x}_i^\top(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\bigg\|_{\infty,\infty}}_{I_3} \\ &+ \underbrace{\bigg\|(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\bigg\|_{\infty,\infty}}_{I_4}.\end{aligned} \tag{B.21}$$

In the following proof, let $\mathbf{C}^{(i)} = (\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top\mathbf{x}_i\boldsymbol{\epsilon}_i^\top$, $\mathbf{D}^{(i)} = \boldsymbol{\epsilon}_i\mathbf{x}_i^\top(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})$ and $\mathbf{E}^{(i)} = (\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top(\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})$. For term I_1, by Lemma D.6, we have, with probability at least $1 - C''/m$

$$\bigg\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top - \boldsymbol{\Sigma}^*\bigg\|_{\infty,\infty} \le C'\nu\sqrt{\frac{\log m}{n}}. \tag{B.22}$$

For term I_2, following the proof procedure for term I_1 in Part I, consider each element in the matrix, i.e., $C_{jk}^{(i)}$, we can easily have

$$\begin{aligned}\Big\|C_{jk}^{(i)}\Big\|_{\psi_1} &= 2\sqrt{C_1C_2\nu\tau}\sqrt{C_3\sum_{\ell=1}^{d}(W_{\ell k}^* - W_{\ell k})^2} \\ &\le 2\sqrt{\nu\tau}\sqrt{C_1C_2C_3}\big\|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*\big\|_F \le 2\sqrt{\nu\tau}\sqrt{C_1C_2C_3}\cdot r\end{aligned}$$

where the last inequality is due to $\|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*\|_F \le r$. Similarly by Bernstein-type inequality in Theorem D.4, we have

$$\mathbb{P}\bigg\{\bigg|\frac{1}{n}\sum_{i=1}^{n}C_{jk}^{(i)}\bigg| \ge t\bigg\} \le \exp\bigg(\frac{C_4t^2n}{4C_1C_2C_3\nu\tau r^2}\bigg).$$

Apply union bound to all possible pairs of $j \in [m]$, $k \in [m]$, we get

$$\mathbb{P}\bigg\{\bigg\|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top\mathbf{x}_i\boldsymbol{\epsilon}_i^\top\bigg\|_{\infty,\infty} \ge t\bigg\} \le m^2\exp\bigg(\frac{C_4t^2n}{4C_1C_2C_3\nu\tau r^2}\bigg).$$

Choose $t = r\sqrt{\nu\tau}C_9\sqrt{2\log m/n}$ and $C_9 = 2\sqrt{C_1C_2C_3/C_4}$, then with probability at least $1-1/m$ we have that

$$\bigg\|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top \mathbf{x}_i\boldsymbol{\epsilon}_i^\top\bigg\|_{\infty,\infty} \le r\sqrt{\nu\tau}C_9\sqrt{\frac{\log m}{n}}. \tag{B.23}$$

For term I_3, since $\mathbf{D}^{(i)} = \mathbf{C}^{(i)\top}$, it holds the same conclusion for term I_3 that with probability at least $1-1/m$ we have

$$\bigg\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i\mathbf{x}_i^\top(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\bigg\|_{\infty,\infty} \le r\sqrt{\nu\tau}C_9\sqrt{\frac{\log m}{n}}. \tag{B.24}$$

For term I_4, denote $\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma}$ as $\mathbf{H}$, $\big(\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\big)$ as $\boldsymbol{\Delta}$, note that

$$\|E_{jk}^{(i)}\|_{\psi_1} = \bigg\|\sum_{\ell=1}^{d}\sum_{u=1}^{d}H_{j\ell}\Delta_{\ell u}H_{uk}\bigg\|_{\psi_1} = \bigg\|\sum_{\ell=1}^{d}\sum_{u=1}^{d}H_{j\ell}\big[x_{i\ell}x_{iu} - (\Sigma_X^*)_{iu}\big]H_{uk}\bigg\|_{\psi_1}. \tag{B.25}$$

Since from Assumption 4.2 we have $\max_{ij}\|x_{ij}\|_{\psi_2} \le C_5\sqrt{\tau}$ we have

$$\begin{aligned}\|E_{jk}^{(i)}\|_{\psi_1} &\le 2\bigg\|\sum_{u=1}^{d}H_{uk}x_{iu}\sum_{\ell=1}^{d}H_{j\ell}x_{i\ell}\bigg\|_{\psi_1} \le 4\bigg\|\sum_{u=1}^{d}H_{uk}x_{iu}\bigg\|_{\psi_2}\cdot\bigg\|\sum_{\ell=1}^{d}H_{j\ell}x_{i\ell}\bigg\|_{\psi_2}\\ &\le 4C_5^2\tau\sqrt{C_6\sum_{u=1}^{d}H_{uk}^2}\sqrt{C_6\sum_{\ell=1}^{d}H_{j\ell}^2} \le 4C_5^2C_6\tau\big\|\mathbf{H}\big\|_F^2,\end{aligned} \tag{B.26}$$

where the first inequality follows from Lemma D.3, the second inequality holds due to Lemma D.2 and the third inequality follows from Lemma D.1. Note that $\big\|\mathbf{H}\big\|_F \le \big\|\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma}\big\|_F \le r$, (B.16) can be further bounded by

$$\|E_{jk}^{(i)}\|_{\psi_1} \le 4C_5^2C_6\tau r^2.$$

By Bernstein-type inequality in Theorem D.4, we have

$$\mathbb{P}\bigg(\bigg|\frac{1}{n}\sum_{i=1}^{n}E_{jk}^{(i)}\bigg| > t\bigg) \le \exp\Big(-\frac{C_{10}nt^2}{16C_5^4C_6^2\tau^2r^4}\Big).$$

Applying union bound to all possible pairs of $j \in [m]$, $k \in [m]$, we get

$$\mathbb{P}\bigg(\bigg\|(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\bigg\|_{\infty,\infty} > t\bigg) \le m^2\cdot\exp\Big(-\frac{C_{10}nt^2}{16C_5^4C_6^2\tau^2r^4}\Big).$$

Choose $t = 4\tau r^2\sqrt{C_5^2C_6/C_{10}}\sqrt{3\log m/n}$ and $C_{11} = 4\sqrt{C_5^2C_6/C_{10}}$, with probability at least $1-1/m$ we have

$$\bigg\|(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})^\top\bigg(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\bigg)(\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma})\bigg\|_{\infty,\infty} \le C_{11}\tau r^2\sqrt{\frac{\log m}{n}}. \tag{B.27}$$

Note that from lemma conditions we have $r \le \sqrt{\nu/\tau}$, submit (B.22), (B.23), (B.24) and (B.27) into (B.21) and apply union bound, we have with probability at least $1-(C''+3)/m$ that

$$\big\|\nabla_2 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}) - \nabla_2 f_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega})\big\|_{\infty,\infty} \le C'\nu\sqrt{\frac{\log m}{n}} \le C'M\sqrt{\frac{\log m}{n}},$$

where the last inequality follows from the fact that $\nu = \|\boldsymbol{\Omega}^*\|_2 \le \|\boldsymbol{\Omega}^*\|_\infty = M$. This completes the proof. $\square$

B.5 Proof of Lemma A.5

In order to prove Lemma A.5, we need the following auxiliary lemma.

Lemma B.1 ((Nesterov, 2004))**.** Under Assumptions 4.1 and 4.2, let $\eta = 2/(L_1 + \mu_1)$, suppose $\boldsymbol{\Gamma}^+$ is obtained by the following gradient descent update form

$$\boldsymbol{\Gamma}^+ = \boldsymbol{\Gamma} - \eta \nabla_1 f\big(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^*\big).$$

We have

$$\|\boldsymbol{\Gamma}^+ - \boldsymbol{\Gamma}^*\|_F \leq \frac{L_1 - \mu_1}{L_1 + \mu_1} \big\|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*\big\|_F. \tag{B.28}$$

Proof of Lemma A.5. For notation simplicity, let $\boldsymbol{\Gamma}^+$ stands for $\overline{\boldsymbol{\Gamma}}^{(t+0.5)}$, $\boldsymbol{\Gamma}$ stands for $\boldsymbol{\Gamma}^{(t)}$ and $\boldsymbol{\Omega}$ stands for $\boldsymbol{\Omega}^{(t)}$.

$$\begin{aligned}\big\|\boldsymbol{\Gamma}^+ - \boldsymbol{\Gamma}^*\big\|_F &= \big\|\boldsymbol{\Gamma} - \eta_1 \nabla_1 f\big(\boldsymbol{\Gamma}, \boldsymbol{\Omega}\big) - \boldsymbol{\Gamma}^*\big\|_F \\ &\leq \big\|\boldsymbol{\Gamma} - \eta_1 \nabla_1 f\big(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^*\big) - \boldsymbol{\Gamma}^*\big\|_F + \eta_1 \big\|\nabla_1 f\big(\boldsymbol{\Gamma}, \boldsymbol{\Omega}\big) - \nabla_1 f\big(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^*\big)\big\|_F, \end{aligned} \tag{B.29}$$

where the inequality holds due to triangle inequality. Submit the conclusion (B.28) in Lemma B.1 into the above equality, we obtain

$$\big\|\boldsymbol{\Gamma}^+ - \boldsymbol{\Gamma}^*\big\|_F \leq \frac{L_1 - \mu_1}{L_1 + \mu_1} \big\|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*\big\|_F + \frac{2\gamma_1}{L_1 + \mu_1} \cdot \big\|\boldsymbol{\Omega} - \boldsymbol{\Omega}^*\big\|_F, \tag{B.30}$$

where the last term on the right side of the above inequality follows from Lemma A.3, in which we obtain $\|\nabla_1 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega}^*) - \nabla_1 f(\boldsymbol{\Gamma}, \boldsymbol{\Omega})\|_F \leq \gamma_1 \cdot \|\boldsymbol{\Omega}^* - \boldsymbol{\Omega}\|_F$. By submitting the definition of L_1, μ_1, η_1 and γ_1 back into (B.30) we complete the proof. □

B.6 Proof of Lemma A.6

We omit the proof since it is similar to the proof of Lemma A.5.

B.7 Proof of Lemma A.8

Proof of Lemma A.8. Consider $\boldsymbol{\Gamma}^{\text{init}}$ computed in Algorithm 3, in fact, each row of $\boldsymbol{\Gamma}^{\text{init}}$ is equal to $[\boldsymbol{\Gamma}^{\text{init}}]_{i*} = \mathcal{ST}\big((\mathbf{X}^\top \mathbf{X} + \epsilon_\Gamma \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}_i, \lambda_\Gamma\big)$, which can be verified as the closed form solution for the following optimization problem:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1, \quad s.t. \quad \|\boldsymbol{\beta} - (\mathbf{X}^\top \mathbf{X} + \epsilon_\Gamma \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}_i\|_\infty \leq \lambda_\Gamma.$$

This is exactly the form of an elementary estimator for high-dimensional linear regression. By Corollary 1 in Yang et al. (2014a) we immediately obtain the the conclusion.

□

B.8 Proof of Lemma A.9

In order to prove Lemma A.9, we need the following auxiliary lemma.

Lemma B.2. Under Assumptions 4.1 and 4.2, if $\big\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*\big\|_F \leq \sqrt{\nu/\tau}$, with probability at least $1 - c_0/m$, we have

$$\|\mathbf{S} - {\boldsymbol{\Omega}^*}^{-1}\|_{\infty,\infty} \leq C\nu \sqrt{\frac{\log m}{n}} + C' \nu^{\frac{1}{2}} \tau^{\frac{5}{6}} d^{\frac{2}{3}} {s_1^*}^{\frac{2}{3}} \cdot \left(\frac{\log dm}{n}\right)^{\frac{5}{6}}.$$

Proof of Lemma A.9. This proof this inspired by Yang et al. (2014b). Consider $\boldsymbol{\Omega}^{\text{init}}$ computed in Algorithm 3, it can be verified that that $\boldsymbol{\Omega}^{\text{init}}$ is the closed form solution for the following optimization problem.

$$\min_{\boldsymbol{\Omega}} \|\boldsymbol{\Omega}\|_{1,1}, \quad s.t. \quad \|\boldsymbol{\Omega} - (\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}\|_{\infty,\infty} \leq \lambda_\Omega,$$

where $\mathbf{S} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma})$ is the sample covariance matrix. Let $\widehat{\boldsymbol{\Omega}}$ denote that solution for the above optimization problem. Following the similar proof as in Theorem 1 in Yang et al. (2014b), we can also show that

$$\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_F \leq 4\sqrt{ms_2^*}\lambda_\Omega, \text{ for } \lambda_\Omega \geq \|\boldsymbol{\Omega}^* - (\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}\|_{\infty,\infty}. \tag{B.31}$$

The remaining task is to find the upper bound for $\|\boldsymbol{\Omega}^* - (\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}\|_{\infty,\infty}$.

$$\begin{aligned}
\|\boldsymbol{\Omega}^* - (\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}\|_{\infty,\infty} &\leq \big\|(\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}((\mathbf{S} + \epsilon_\Gamma \mathbf{I})\boldsymbol{\Omega}^* - \mathbf{I})\big\|_{\infty,\infty} \\
&\leq \|(\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}\|_\infty \cdot \Big\|\boldsymbol{\Omega}^*\big(\mathbf{S} + \epsilon_\Gamma \mathbf{I}) - \boldsymbol{\Omega}^{*-1}\big)\Big\|_{\infty,\infty} \\
&\leq \|(\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}\|_\infty \cdot \|\boldsymbol{\Omega}^*\|_\infty \cdot \Big\|\big(\mathbf{S} + \epsilon_\Gamma \mathbf{I}) - \boldsymbol{\Omega}^{*-1}\big)\Big\|_{\infty,\infty} \\
&\leq M\|(\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}\|_\infty \cdot \big\|(\mathbf{S} + \epsilon_\Gamma \mathbf{I}) - \boldsymbol{\Omega}^{*-1}\big\|_{\infty,\infty},
\end{aligned}$$

where the last inequality follows from the assumption that $\|\boldsymbol{\Omega}^*\|_\infty = \|\boldsymbol{\Omega}^*\|_1 \leq M$. Following the similar proof and suppose the same condition in Corollary 1 of Yang et al. (2014b) also holds, further we have

$$\begin{aligned}
\|\boldsymbol{\Omega}^* - (\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}\|_{\infty,\infty} &\leq CM\big\|(\mathbf{S} + \epsilon_\Gamma \mathbf{I}) - \boldsymbol{\Omega}^{*-1}\big\|_{\infty,\infty} \\
&\leq CM\big(\|\mathbf{S} - \boldsymbol{\Omega}^{*-1}\|_{\infty,\infty} + \epsilon_\Gamma\big),
\end{aligned}$$

By combining Lemma B.2 with the above result and choose ϵ_Γ as the upper bound for $\|\mathbf{S} - \boldsymbol{\Omega}^{*-1}\|_{\infty,\infty}$ we have

$$\|\boldsymbol{\Omega}^* - (\mathbf{S} + \epsilon_\Gamma \mathbf{I})^{-1}\|_{\infty,\infty} \leq C'M\nu\sqrt{\frac{\log m}{n}} + C''\nu^{\frac{1}{2}}\tau^{\frac{5}{6}}(ds_1^*)^{\frac{2}{3}} \cdot \left(\frac{\log dm}{n}\right)^{\frac{5}{6}} := \lambda_\Omega.$$

Submit the value for λ_Ω back into (B.31) we have

$$\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*\|_F \leq C'M\nu\sqrt{\frac{ms_2^* \cdot \log m}{n}} + C''\nu^{\frac{1}{2}}\tau^{\frac{5}{6}}(ms_2^*)^{\frac{1}{2}}(ds_1^*)^{\frac{2}{3}} \cdot \left(\frac{\log dm}{n}\right)^{\frac{5}{6}}.$$

The proof is completed. □

C Proof of Auxiliary Lemmas in Section B

C.1 Proof of Lemma B.2

Proof. Let us denote $\boldsymbol{\Sigma}^* = \boldsymbol{\Omega}^{*-1}$. Since we have

$$\begin{aligned}
\mathbf{S} &= \frac{1}{n}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Gamma}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Gamma}}) \\
&= \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}^* + \mathbf{X}\boldsymbol{\Gamma}^* - \mathbf{X}\widehat{\boldsymbol{\Gamma}})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}^* + \mathbf{X}\boldsymbol{\Gamma}^* - \mathbf{X}\widehat{\boldsymbol{\Gamma}}) \\
&= \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}^*)^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}^*) + \frac{1}{n}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)^\top\mathbf{X}^\top\mathbf{X}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*) + \frac{2}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}^*)^\top\mathbf{X}(\boldsymbol{\Gamma}^* - \widehat{\boldsymbol{\Gamma}}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\|\mathbf{S} - \boldsymbol{\Sigma}^*\|_{\infty,\infty} \\
&\leq \underbrace{\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\infty,\infty}}_{I_1} + \underbrace{\left\|\frac{2}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\mathbf{x}_i^\top(\boldsymbol{\Gamma}^* - \widehat{\boldsymbol{\Gamma}})\right\|_{\infty,\infty}}_{I_2} + \underbrace{\left\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)^\top\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^*\right)(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)\right\|_{\infty,\infty}}_{I_3} \\
&\quad + \underbrace{\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)^\top\boldsymbol{\Sigma}_X^*(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)\|_{\infty,\infty}}_{I_4}.
\end{aligned} \tag{C.1}$$

For term I_1, by Lemma D.6, we have

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\infty,\infty} \leq C_1 \lambda_{\max}(\boldsymbol{\Sigma}^*) \sqrt{\frac{\log m}{n}} = C_1 \nu \sqrt{\frac{\log m}{n}}.$$

For term I_2, using similar techniques as we do for term I_3 in (B.21) we have

$$\bigg\| \frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\epsilon}_i \mathbf{x}_i^\top (\boldsymbol{\Gamma}^* - \widehat{\boldsymbol{\Gamma}}) \bigg\|_{\infty,\infty} \leq C_2 \sqrt{\nu\tau} \cdot \|\boldsymbol{\Gamma}^* - \widehat{\boldsymbol{\Gamma}}\|_F \cdot \sqrt{\frac{\log m}{n}} \leq C_2' \underbrace{\nu^{\frac{1}{2}} \tau^{\frac{5}{6}} d^{\frac{2}{3}} s_1^{*\frac{2}{3}} \cdot \left(\frac{\log dm}{n}\right)^{\frac{5}{6}}}_{B}.$$

For term I_3, using similar technique as we do for term I_4 in (B.21) and also since $\|\boldsymbol{\Gamma}^* - \widehat{\boldsymbol{\Gamma}}\|_F \leq \sqrt{\nu/\tau}$ we have

$$\begin{aligned}
\bigg\| (\boldsymbol{\Gamma}^* - \widehat{\boldsymbol{\Gamma}})^\top \bigg(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top - \boldsymbol{\Sigma}_X^* \bigg) (\boldsymbol{\Gamma}^* - \widehat{\boldsymbol{\Gamma}}) \bigg\|_{\infty,\infty} &\leq C_3 \tau \|\boldsymbol{\Gamma}^* - \widehat{\boldsymbol{\Gamma}}\|_F^2 \cdot \sqrt{\frac{\log m}{n}} \\
&\leq C_3' \sqrt{\nu\tau} \cdot \|\boldsymbol{\Gamma}^* - \widehat{\boldsymbol{\Gamma}}\|_F \cdot \sqrt{\frac{\log m}{n}} \leq C_3 B.
\end{aligned}$$

For term I_4, we also have

$$\begin{aligned}
\|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)^\top \boldsymbol{\Sigma}_X^* (\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)\|_{\infty,\infty} &\leq \|(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)^\top \boldsymbol{\Sigma}_X^* (\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*)\|_2 \\
&\leq \|\boldsymbol{\Sigma}_X^*\|_2 \sqrt{\nu/\tau} \cdot \|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}^*\|_F \\
&\leq C_4 B,
\end{aligned}$$

where the first and second inequalities hold due to the matrix norm inequalities, and the third inequality follows from Assumption 4.2 and also the conclusion from Lemma A.8. Combine the results for term I_1, I_2, I_3 and I_4, we obtain

$$\|\mathbf{S} - \boldsymbol{\Sigma}^*\|_{\infty,\infty} \leq C_1 \nu \sqrt{\frac{\log m}{n}} + C_5 \nu^{\frac{1}{2}} \tau^{\frac{5}{6}} d^{\frac{2}{3}} s_1^{*\frac{2}{3}} \cdot \left(\frac{\log dm}{n}\right)^{\frac{5}{6}}. \tag{C.2}$$

This completes the proof. □

D Additional Auxiliary Lemmas

Lemma D.1 (Rotation invariance (Vershynin, 2010)). For a set of independent centered sub-Gaussian random variables X_i, $\sum_i a_i \|X_i\|_{\nu_2}^2$ is also a centered sub-Gaussian random variable, and further, we have

$$\Big\| \sum_i a_i X_i \Big\|_{\psi_2}^2 \leq C \sum_i a_i^2 \|X_i\|_{\psi_2}^2,$$

where C is an absolute constant.

Lemma D.2 (Product Property (Vershynin, 2010)). For any two sub-Gaussian random variables X and Y, we have

$$\|XY\|_{\psi_1} \leq 2\|X\|_{\psi_2} \cdot \|Y\|_{\psi_2}.$$

Lemma D.3 (Centering (Vershynin, 2010)). For any sub-Exponential random variables X, we have

$$\|X - \mathbb{E}X\|_{\psi_1} \leq 2\|X\|_{\psi_1}.$$

Theorem D.4 (Proposition 5.16 in (Vershynin, 2010)). Let $X_1, X_2, \ldots, X_n$ be independent centered sub-exponential random variables, and let $K = \max_i \|X_i\|_{\psi_1}$. Then for every $a = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ and for every $t > 0$, we have

$$\mathbb{P}\bigg(\bigg| \sum_{i=1}^{n} a_i X_i \bigg| > t \bigg) \leq 2 \exp\Big[-C \min\Big(\frac{t^2}{K^2 \|a\|_2^2}, \frac{t}{K\|a\|_\infty} \Big) \Big],$$

where $C > 0$ is a constant.

Lemma D.5 ((Vershynin, 2010))**.** Suppose $S \subseteq \mathbb{R}^d$ is an index set with $|S| = s$, we have with probability at least $1 - 1/n^2$ that

$$\|\widehat{\boldsymbol{\Sigma}}_{SS} - \boldsymbol{\Sigma}_{SS}\|_2 \le C\lambda_{\max}(\boldsymbol{\Sigma}_{SS})\sqrt{\frac{s}{n}},$$

where C is some universal constant.

Lemma D.6 ((Loh & Wainwright, 2013))**.** Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. sub-Gaussian random vectors in $\mathbb{R}^d$, and $\boldsymbol{\Sigma}^* = \mathbb{E}[1/n \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top]$. We have with probability at least $1 - C/d$ that

$$\bigg\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top - \boldsymbol{\Sigma}^*\bigg\|_{\infty,\infty} \le C\lambda_{\max}(\boldsymbol{\Sigma}^*)\sqrt{\frac{\log d}{n}},$$

where C is an absolute constant.

Lemma D.7. (Zhou, 2009) For any sub-Gaussian random design $\mathbf{X} \in \mathbb{R}^{n\times d}$ with i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma})$ rows, there are strictly positive constants (κ_1, κ_2), depending only on the positive definite matrix $\boldsymbol{\Sigma}$, such that for any $\mathbf{v} \in \mathbb{R}^d$, we have

$$\frac{\|\mathbf{X}\mathbf{v}\|_2^2}{n} \ge \kappa_1\|\mathbf{v}\|_2^2 - \kappa_2\frac{\log d}{n}\|\mathbf{v}\|_1^2$$

holds with probability at least $1 - C'\exp(-Cn)$, where C, C' are positive constants.

E Additional Experimental Materials

Figures 6, 7, 8 show the gene networks recovered by Alt-NCD, MRCE and Capme respectively. Figure 9 shows the cell cycle Saccharomyces cerevisiae pathway from KEGG database. It shows that our method can discover more meaningful interactions.

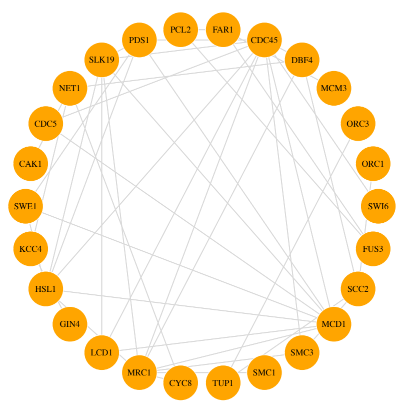

Figure 6. Gene network recovered by Alt-NCD for the 92 genes on the cell-cycle yeast pathway.

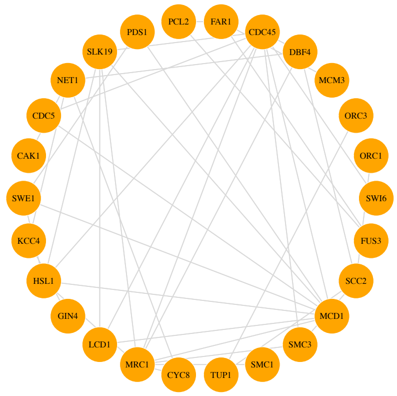

Figure 7. Gene network recovered by MRCE for the 92 genes on the cell-cycle yeast pathway.

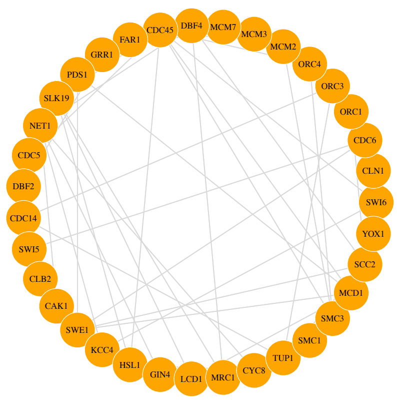

Figure 8. Gene network recovered by Capme for the 92 genes on the cell-cycle yeast pathway.

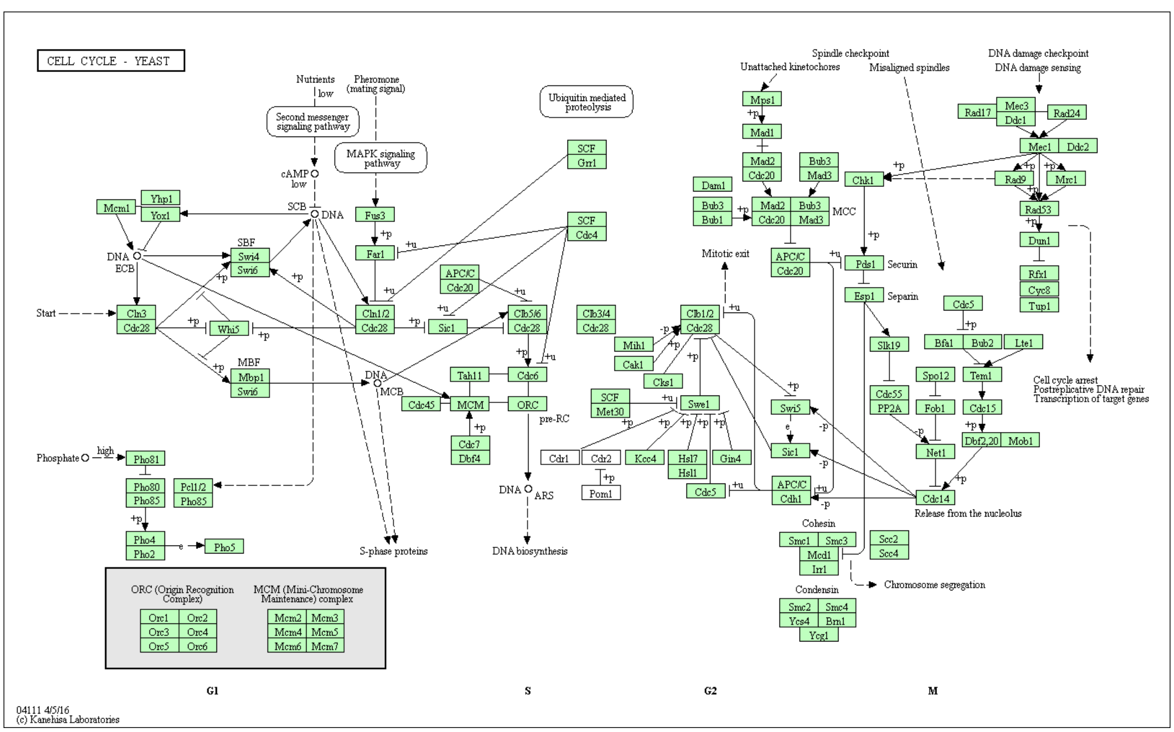

Figure 9. Cell cycle Saccharomyces cerevisiae pathway from KEGG database