8. Appendix

We prove Theorem 1 in this section. Since the proof is technical and lengthy, for the clarity of presentation, we organize the proof as follows. To begin with, in Section 8.1, we review two standard concentration inequalities, the Chernoff inequality and the Hoeffding inequality, which will be used to prove some technical lemmas. We then present and prove these technical lemmas in Section 8.2. These technical lemmas are subsequently used to validate some auxiliary results, which are presented in Section 8.3. Finally, we prove Theorem 1 based on these auxiliary results.

8.1. Concentration Inequalities

Lemma 1 (Hoeffding Inequality)**.** *Let $X_1, X_2, \cdots, X_n$ be n i.i.d. random variables drawn from the distribution $\mathcal{D}$, with $0 \leq X_i \leq a$, $\forall i \in \{1, 2, \cdots, n\}$. Let $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for any $t > 0$,*

$$P(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq t) \leq 2\exp\left(-\frac{2nt^2}{a^2}\right).$$

Lemma 2 (Chernoff Inequality)**.** *Let $X_1, X_2, \cdots, X_n$ be n random variables and let $X := \sum_{i=1}^{n} X_i$. Then, for any $t > 0$,*

$$P(X \geq \epsilon) \leq \exp(-t\epsilon)\mathbb{E}\left[\exp\left(\sum_{i=1}^{n} tX_i\right)\right]. \tag{14}$$

Furthermore, if X_i's are independent, then

$$P(X \geq \epsilon) \leq \min_{t>0} \exp(-t\epsilon)\prod_{i=1}^{n}\mathbb{E}\left[\exp(tX_i)\right]. \tag{15}$$

8.2. Technical Lemmas

We use $\|\cdot\|_{\max}$ to represent the max norm of a matrix, which is equal to the maximum of the absolute value of all the elements in the matrix.

Lemma 3. *Let $\mathbb{X}$ be given. Suppose that $0 < \max_{i,i' \in \{1,2,\cdots,n\}} \|\mathbf{x}_i \mathbf{x}_{i'}^\top\|_{\max} < \epsilon^2$. Then,*

$$\mathrm{P}\left(\max_{j \neq j', j \neq j', j, j' \in \{1,2,\cdots,p\}} |\mathbb{E}_{\mathbb{X}}[X_j X_{j'}] - \mathbb{E}[X_j X_{j'}]| \geq \epsilon^2 \sqrt{\frac{\log p}{n}}\right) \leq 2\exp(-2\log p).$$

Proof. Since $0 < \max_{i,i' \in \{1,2,\cdots,n\}} \|\mathbf{x}_i \mathbf{x}_{i'}^\top\|_{\max} < \epsilon^2$, we let $a = \epsilon^2$ and $t = \epsilon^2\sqrt{\frac{\log p}{n}}$ in Lemma 1 to yield the result. □

Lemma 4. *Let $\mathbb{X}$ be given. Suppose that $0 < \max_{i \in \{1,2,\cdots,n\}} \|\mathbf{x}_i\|_\infty < \epsilon$. Then,*

$$\mathrm{P}\left(\max_{j \in \{1,2,\cdots,p\}} |\mathbb{E}_{\mathbb{X}}[X_j] - \mathbb{E}[X_j]| \geq \epsilon \sqrt{\frac{\log p}{n}}\right) \leq 2\exp(-2\log p).$$

Proof. Since $0 < \max_{i \in \{1,2,\cdots,n\}} \|\mathbf{x}_i\|_\infty < \epsilon$, we let $a = \epsilon$ and $t = \epsilon\sqrt{\frac{\log p}{n}}$ in Lemma 1 to yield the result. □

Lemma 5. *Let $\mathbb{X}$ be given. Suppose that $0 < \max_{i \in \{1,2,\cdots,n\}} \|\mathbf{x}_i\|_\infty < \epsilon$. Then,*

$$\mathrm{P}\left(\max_{j, j' \in \{1,2,\cdots,p\}} |\mathbb{E}_{\mathbb{X}}[\mathbb{E}[X_j X_{j'} \mid \mathbf{X}_{-j}]] - \mathbb{E}[\mathbb{E}[X_j X_{j'} \mid \mathbf{X}_{-j}]]| \geq C_3\epsilon \sqrt{\frac{\log p}{n}}\right) \leq 2\exp(-2\log p).$$

Proof. Since $0 < \max_{i \in \{1,2,\cdots,n\}} \|\mathbf{x}_i\|_\infty < \epsilon$ and $\mathbb{E}[X_j | \mathbf{x}_{i,-j}] \leq C_3$ by Assumption 2, we have that $0 < \mathbb{E}[X_j X_{j'} | \mathbf{x}_{i,-j}] \leq C_3\epsilon$. Therefore, we let $a = C_3\epsilon$ and $t = C_3\epsilon\sqrt{\frac{\log p}{n}}$ in Lemma 1 to yield the result. □

Remark

The subtlety of the definitions of C_3 and C_4 in Assumption 2, as well as the notion of ϵ in Lemma 3, Lemma 4, and Lemma 5 should be noted. Formally, the n data points $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ in $\mathbb{X}$ can be viewed as assignments to the corresponding random variables $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(n)}$ following the PSQR parameterized by $\mathbf{\Theta}^*$. In Assumption 2, we are interested in a set $\mathcal{X} \subseteq \mathbb{N}^p$, such that $\forall i \in \{1, 2, \cdots, n\}$ and $\forall j \in \{1, 2, \cdots, p\}$,

$$\max_{\mathbf{X}^{(i)} \in \mathcal{X}} \mathbb{E}\left[X_j \mid \mathbf{X}^{(i)}_{-j}\right] \leq C_3 \quad \text{and} \quad \max_{\mathbf{X}^{(i)} \in \mathcal{X}} \left|\lambda^*_{ij} - \mathbb{E}\left[X_j \mid \mathbf{X}^{(i)}_{-j}\right]\right| \leq C_4.$$

In Lemma 3, Lemma 4, and Lemma 5, we are interested in a set $\mathcal{X} \subseteq \mathbb{N}^p$, such that $\forall i, i' \in \{1, 2, \cdots, n\}$, where $i \neq i'$,

$$0 < \max_{\mathbf{X}^{(i)}, \mathbf{X}^{(i')} \in \mathcal{X}} \|\mathbf{X}^{(i)} \mathbf{X}^{(i')\top}\|_{\max} < \epsilon^2 \quad \text{and} \quad 0 < \max_{\mathbf{X}^{(i)} \in \mathcal{X}} \|\mathbf{X}^{(i)}\|_\infty < \epsilon.$$

Also, implicitly, we have that $\mathbf{x}_i \in \mathcal{X}$, $\forall i \in \{1, 2, \cdots, n\}$.

Lemma 6. *Let $\mathbb{X}$ be given. Then,*

$$\mathrm{P}\left(\max_{j \in \{1,2,\cdots,p\}} |\mathbb{E}_{\mathbb{X}}[\mathbb{E}[X_j \mid \mathbf{X}_{-j}]] - \mathbb{E}[\mathbb{E}[X_j \mid \mathbf{X}_{-j}]]| \geq C_3 \sqrt{\frac{\log p}{n}}\right) \leq 2\exp(-2\log p).$$

Proof. Since $\mathbb{E}[X_j \mid \mathbf{x}_{i,-j}] \leq C_3$ by Assumption 2, we have that $0 < \mathbb{E}[X_j \mid \mathbf{x}_{i,-j}] \leq C_3$. Therefore, we let $a = C_3$ and $t = C_3\sqrt{\frac{\log p}{n}}$ in Lemma 1 to yield the result. □

Lemma 7. *Let $\mathbf{X}$ be a random vector drawn from a PSQR distribution parameterized by $\mathbf{\Theta}^*$. Suppose that $\{\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n\}^\top$ is the set of n i.i.d. samples of $\mathbf{X}$. Given $j \in \{1, 2, \cdots, p\}$, $\epsilon_1 := 3\log p + \log n$, and $\epsilon_2 := C_1 + \sqrt{\frac{2\log p}{n}}$,*

$$\mathrm{P}\left(X_j \geq \epsilon_1\right) \leq \exp(C_1 + C_2/2 - \epsilon_1), \quad \textit{and} \quad \mathrm{P}\left(\frac{1}{n}\sum_{i=1}^n x_{ij} \geq \epsilon_2\right) \leq \exp\left[-\frac{n(\epsilon_2 - C_1)^2}{2C_2}\right].$$

Proof. We start with proving the first inequality. To this end, consider the following equation due to Taylor expansion:

$$\begin{aligned}\log \mathbb{E}\left[\exp(X_j)\right] =& B(\mathbf{\Theta}^*, \mathbf{0} + \mathbf{e}_j) - B(\mathbf{\Theta}^*, \mathbf{0}) = \boldsymbol{\nabla}^\top B(\mathbf{\Theta}^*, \mathbf{0})\mathbf{e}_j + \frac{1}{2}\mathbf{e}_j^\top \boldsymbol{\nabla}^2 B(\mathbf{\Theta}^*, k\mathbf{e}_j)\mathbf{e}_j \\ =& \mathbb{E}[X_j] + \frac{1}{2}\frac{\partial^2}{\partial b_j^2} B(\mathbf{\Theta}^*, \mathbf{0} + k\mathbf{e}_j) \leq C_1 + C_2/2,\end{aligned} \tag{16}$$

where $k \in [0, 1]$, $\mathbf{e}_j$ is a vector whose j^{th} component is one and zeros elsewhere, and the last inequality is due to Assumption 1. Then, let $t = 1$ and $X = X_j$ in Lemma 2,

$$\mathrm{P}\left(X_j \geq \epsilon_1\right) = \exp(-\epsilon_1)\mathbb{E}\left[\exp(X_j)\right] \leq \exp(C_1 + C_2/2 - \epsilon_1).$$

Now, we prove the second bound. For any $0 < a < 1$ and some $k \in [0, 1]$, with Taylor expansion,

$$\begin{aligned}\log \mathbb{E}\left[\exp(aX_i)\right] =& B(\mathbf{\Theta}^*, \mathbf{0} + a\mathbf{e}_j) - B(\mathbf{\Theta}^*, \mathbf{0}) = a\boldsymbol{\nabla}^\top B(\mathbf{\Theta}^*, \mathbf{0})\mathbf{e}_j + \frac{a^2}{2}\mathbf{e}_j^\top \boldsymbol{\nabla}^2 B(\mathbf{\Theta}^*, \mathbf{0} + ak\mathbf{e}_j)\mathbf{e}_j \\ =& a\mathbb{E}(X_j) + \frac{a^2}{2}\frac{\partial^2}{\partial b_j^2} B(\mathbf{\Theta}^*, \mathbf{0} + ak\mathbf{e}_j) \leq aC_1 + \frac{a^2}{2}C_2,\end{aligned} \tag{17}$$

where the last inequality is due to Assumption 1. Then, following the proof technique above, we have

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon_2\right) = \mathrm{P}\left(\sum_{i=1}^n X_i \geq n\epsilon_2\right) \leq \min_{t>0} \exp(-tn\epsilon_2) \prod_{i=1}^n \mathbb{E}\left[\exp\left(tX_i\right)\right]$$

$$\leq \min_{t>0} \exp(-tn\epsilon_2) \prod_{i=1}^{n} \exp\left(C_1 t + \frac{C_2}{2} t^2\right)$$
$$= \min_{t>0} \exp\left[(C_1 - \epsilon_2)nt + \frac{nC_2}{2} t^2\right] \leq \exp\left[-\frac{n(\epsilon_2 - C_1)^2}{2C_2}\right],$$

where the minimum is obtained when $t = \frac{\epsilon_2 - C_1}{C_2}$, and we have used the fact that $\epsilon_2 > C_1$. □

8.3. Auxiliary Results

Lemma 8. *Let $r := 4C_5\lambda$. Then with probability of at least $1 - \left((\exp(C_1 + C_2/2) + 8)\, p^{-2} + p^{-1/C_2}\right)$, the following two inequalities simultaneously hold:*

$$\|\nabla F(\mathbf{\Theta}^*)\|_\infty \leq 2\left[C_3(3\log p + \log n) + (3\log p + \log n)^2\right]\sqrt{\frac{\log p}{n}} + 2C_4\left(C_1 + \sqrt{\frac{2\log p}{n}}\right), \tag{18}$$

$$\|\tilde{\mathbf{\Theta}}_S - \mathbf{\Theta}^*_S\|_\infty \leq r. \tag{19}$$

Proof. We prove (18) and (19) in turn.

PROOF OF (18)

To begin with, we prove (18). By the definition of F in (13), for $j < j'$, the derivative of $F(\mathbf{\Theta}^*)$ is:

$$\frac{\partial F(\mathbf{\Theta}^*)}{\partial \theta_{jj'}} = \frac{1}{n}\sum_{i=1}^{n}\left[-x_{ij'}x_{ij} + \lambda^*_{ij}x_{ij'} - x_{ij}x_{ij'} + \lambda^*_{ij'}x_{ij}\right] = -2\mathbb{E}_{\mathbb{X}}[X_jX_{j'}] + \frac{1}{n}\sum_{i=1}^{n}\lambda^*_{ij}x_{ij'} + \frac{1}{n}\sum_{i=1}^{n}\lambda^*_{ij'}x_{ij}. \tag{20}$$

and

$$\frac{\partial}{\partial \theta_{jj}} F(\mathbf{\Theta}^*) = \frac{1}{n}\sum_{i=1}^{n}\left[-x_{ij} + \lambda^*_{ij}\right] = -\mathbb{E}_{\mathbb{X}}[X_j] + \frac{1}{n}\sum_{i=1}^{n}\lambda^*_{ij}, \tag{21}$$

where $\mathbb{E}_{\mathbb{X}}[X_jX_{j'}] := \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ij'}$ and $\mathbb{E}_{\mathbb{X}}[X_j] := \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ are the expectations of $X_jX_{j'}$ and X_j over the empirical distribution given by the dataset $\mathbb{X}$.

Then, by defining $\mathbb{E}[X_jX_{j'}]$ as the expectation of the multiplication of two components of an multivariate square root Poisson random vector whose distribution is parameterized by $\mathbf{\Theta}^*$, and by Assumption 2, (20) can be controlled via

$$\begin{aligned}
\left|\frac{\partial}{\partial \theta_{jj'}} F(\mathbf{\Theta}^*)\right| =& \left|\frac{1}{n}\sum_{i=1}^{n}\lambda^*_{ij}x_{ij'} - \mathbb{E}[X_jX_{j'}] + \frac{1}{n}\sum_{i=1}^{n}\lambda^*_{ij'}x_{ij} - \mathbb{E}[X_jX_{j'}] + 2\mathbb{E}[X_jX_{j'}] - 2\mathbb{E}_{\mathbb{X}}[X_jX_{j'}]\right| \\
\leq& \left|\frac{1}{n}\sum_{i=1}^{n}\lambda^*_{ij}x_{ij'} - \mathbb{E}[X_jX_{j'}]\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\lambda^*_{ij'}x_{ij} - \mathbb{E}[X_jX_{j'}]\right| + 2|\mathbb{E}_{\mathbb{X}}[X_jX_{j'}] - \mathbb{E}[X_jX_{j'}]| \\
=& \left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[X_j \mid \mathbf{X}_{-j} = \mathbf{x}_{i,-j}] + \lambda^*_{ij} - \mathbb{E}[X_j \mid \mathbf{X}_{-j} = \mathbf{x}_{i,-j}]\right)x_{ij'} - \mathbb{E}[X_jX_{j'}]\right| \\
&+ \left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[X_{j'} \mid \mathbf{X}_{-j'} = \mathbf{x}_{i,-j'}] + \lambda^*_{ij'} - \mathbb{E}[X_{j'} \mid \mathbf{X}_{-j'} = \mathbf{x}_{i,-j'}]\right)x_{ij} - \mathbb{E}[X_jX_{j'}]\right| \\
&+ 2|\mathbb{E}_{\mathbb{X}}[X_jX_{j'}] - \mathbb{E}[X_jX_{j'}]| \\
\leq& \left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[X_j \mid \mathbf{X}_{-j} = \mathbf{x}_{i,-j}]\right)x_{ij'} - \mathbb{E}[X_jX_{j'}]\right| + \frac{1}{n}\sum_{i=1}^{n}\left|\lambda^*_{ij} - \mathbb{E}[X_j \mid \mathbf{X}_{-j} = \mathbf{x}_{i,-j}]\right|x_{ij'} \\
&+ \left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[X_{j'} \mid \mathbf{X}_{-j'} = \mathbf{x}_{i,-j'}]\right)x_{ij} - \mathbb{E}[X_jX_{j'}]\right| + \frac{1}{n}\sum_{i=1}^{n}\left|\lambda^*_{ij'} - \mathbb{E}[X_{j'} \mid \mathbf{X}_{-j'} = \mathbf{x}_{i,-j'}]\right|x_{ij} \\
&+ 2|\mathbb{E}_{\mathbb{X}}[X_jX_{j'}] - \mathbb{E}[X_jX_{j'}]|
\end{aligned}$$

$$
\begin{aligned}
\leq & \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_j \mid \mathbf{X}_{-j} = \mathbf{x}_{i,-j}] x_{ij'} - \mathbb{E}[X_j X_{j'}] \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_{j'} \mid \mathbf{X}_{-j'} = \mathbf{x}_{i,-j'}] x_{ij} - \mathbb{E}[X_j X_{j'}] \right| \\
& + 2|\mathbb{E}_{\mathbb{X}}[X_j X_{j'}] - \mathbb{E}[X_j X_{j'}]| + C_4(\mathbb{E}_{\mathbb{X}}[X_j] + \mathbb{E}_{\mathbb{X}}[X_{j'}]) \\
= & \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_j X_{j'} \mid \mathbf{X}_{-j} = \mathbf{x}_{i,-j}] - \mathbb{E}[X_j X_{j'}] \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_j X_{j'} \mid \mathbf{X}_{-j'} = \mathbf{x}_{i,-j'}] - \mathbb{E}[X_j X_{j'}] \right| \\
& + 2|\mathbb{E}_{\mathbb{X}}[X_j X_{j'}] - \mathbb{E}[X_j X_{j'}]| + C_4(\mathbb{E}_{\mathbb{X}}[X_j] + \mathbb{E}_{\mathbb{X}}[X_{j'}]) \\
= & 2|\mathbb{E}_{\mathbb{X}}[\mathbb{E}[X_j X_{j'} \mid \mathbf{X}_{-j}]] - \mathbb{E}[X_j X_{j'}]| + 2|\mathbb{E}_{\mathbb{X}}[X_j X_{j'}] - \mathbb{E}[X_j X_{j'}]| + C_4(\mathbb{E}_{\mathbb{X}}[X_j] + \mathbb{E}_{\mathbb{X}}[X_{j'}]) \\
= & 2|\mathbb{E}_{\mathbb{X}}[\mathbb{E}[X_j X_{j'} \mid \mathbf{X}_{-j}]] - \mathbb{E}[\mathbb{E}[X_j X_{j'} \mid \mathbf{X}_{-j}]]| + 2|\mathbb{E}_{\mathbb{X}}[X_j X_{j'}] - \mathbb{E}[X_j X_{j'}]| + C_4(\mathbb{E}_{\mathbb{X}}[X_j] + \mathbb{E}_{\mathbb{X}}[X_{j'}]),
\end{aligned}
$$

where we have used the law of total expectation in the last equality.

Similarly, (21) can be controlled via

$$
\begin{aligned}
\left| \frac{\partial}{\partial \theta_{jj}} F(\mathbf{\Theta}^*) \right| = & \left| -\mathbb{E}_{\mathbb{X}}[X_j] + \frac{1}{n} \sum_{i=1}^{n} \lambda_{ij}^* \right| = \left| -\mathbb{E}_{\mathbb{X}}[X_j] + \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}[X_j \mid \mathbf{X}_{-j} = \mathbf{x}_{i,-j}] + \lambda_{ij}^* - \mathbb{E}[X_j \mid \mathbf{X}_{-j} = \mathbf{x}_{i,-j}] \right) \right| \\
= & \left| -\mathbb{E}_{\mathbb{X}}[X_j] + \mathbb{E}[X_j] - \mathbb{E}[X_j] + \mathbb{E}_{\mathbb{X}}[\mathbb{E}[X_j \mid \mathbf{X}_{-j}]] + \frac{1}{n} \sum_{i=1}^{n} \left(\lambda_{ij}^* - \mathbb{E}[X_j \mid \mathbf{X}_{-j} = \mathbf{x}_{i,-j}] \right) \right| \\
\leq & |\mathbb{E}_{\mathbb{X}}[\mathbb{E}[X_j \mid \mathbf{X}_{-j}]] - \mathbb{E}[X_j]| + |\mathbb{E}_{\mathbb{X}}[X_j] - \mathbb{E}[X_j]| + C_4 \\
= & |\mathbb{E}_{\mathbb{X}}[\mathbb{E}[X_j \mid \mathbf{X}_{-j}]] - \mathbb{E}[\mathbb{E}[X_j \mid \mathbf{X}_{-j}]]| + |\mathbb{E}_{\mathbb{X}}[X_j] - \mathbb{E}[X_j]| + C_4.
\end{aligned}
$$

We define four events:

$$
\begin{aligned}
E_1 &:= \left\{ \max_{j \neq j', j, j' \in \{1,2,\cdots,p\}} \left| \frac{\partial}{\partial \theta_{jj'}} F(\mathbf{\Theta}^*) \right| \geq 2(C_3 \epsilon_1 + \epsilon_1^2) \sqrt{\frac{\log p}{n}} + 2C_4 \epsilon_2 \right\}, \\
E_2 &:= \left\{ \max_{j \in \{1,2,\cdots,p\}} \left| \frac{\partial}{\partial \theta_{jj}} F(\mathbf{\Theta}^*) \right| \geq (C_3 + \epsilon_1) \sqrt{\frac{\log p}{n}} + C_4/n \right\}, \\
E_3 &:= \left\{ 0 < \max_{i \in \{1,2,\cdots,n\}} \|\mathbf{x}_i\|_\infty < \epsilon_1 \right\}, \quad \text{and} \quad E_4 := \left\{ 0 < \max_{j \in \{1,2,\cdots,p\}} \mathbb{E}_{\mathbb{X}}[X_j] < \epsilon_2 \right\},
\end{aligned}
$$

where $\epsilon_1 := 3 \log p + \log n$ and $\epsilon_2 := C_1 + \sqrt{\frac{2 \log p}{n}}$ are defined in Lemma 7. By Lemma 3, Lemma 4, Lemma 5 and Lemma 6, it follows that

$$
\mathrm{P}(E_1 \mid E_3, E_4) \leq 4 \exp(-2 \log p) \quad \text{and} \quad \mathrm{P}(E_2 \mid E_3, E_4) \leq 4 \exp(-2 \log p). \tag{22}
$$

Therefore,

$$
\begin{aligned}
\mathrm{P}(E_1 \cup E_2) = & \mathrm{P}(E_1 \cup E_2 \mid E_3, E_4) \mathrm{P}(E_3, E_4) + \mathrm{P}(E_1 \cup E_2 \mid E_3^c, E_4) \mathrm{P}(E_3^c, E_4) \\
& + \mathrm{P}(E_1 \cup E_2 \mid E_3, E_4^c) \mathrm{P}(E_3, E_4^c) + \mathrm{P}(E_1 \cup E_2 \mid E_3^c, E_4^c) \mathrm{P}(E_3^c, E_4^c) \\
\leq & \mathrm{P}(E_1 \mid E_3, E_4) + \mathrm{P}(E_2 \mid E_3, E_4) + \mathrm{P}(E_3^c, E_4) + \mathrm{P}(E_3, E_4^c) + \mathrm{P}(E_3^C, E_4^c) \\
\leq & \mathrm{P}(E_1 \mid E_3, E_4) + \mathrm{P}(E_2 \mid E_3, E_4) + \mathrm{P}(E_3^c) + \mathrm{P}(E_4^c) \\
\leq & 8 \exp(-2 \log p) + \exp(C_1 + C_2/2 - \epsilon_1) np + \exp\left[-\frac{n(\epsilon_2 - C_1)^2}{2C_2} \right] \\
= & 8 \exp(-2 \log p) + \frac{\exp(C_1 + C_2/2)}{p^2} + p^{-\frac{1}{C_2}},
\end{aligned} \tag{23}
$$

where the superscript c over an event represents the complement of that event, and the last inequality is due to (22) and Lemma 7. Also notice that by the definitions of E_1 and E_2,

$$
2(C_3 \epsilon_1 + \epsilon_1^2) \sqrt{\frac{\log p}{n}} + 2C_4 \epsilon_2 > (C_3 + \epsilon_1) \sqrt{\frac{\log p}{n}} + C_4/n.
$$

Therefore, with probability of $1-\mathrm{P}(E_1\cup E_2)\geq 1-\left((\exp\left(C_1+C_2/2\right)+8\right)p^{-2}+p^{-1/C_2}\big)$, neither E_1 nor E_2 occurs, and hence

$$\begin{aligned}\|\boldsymbol{\nabla}F(\boldsymbol{\Theta}^*)\|_\infty \leq & 2(C_3\epsilon_1+\epsilon_1^2)\sqrt{\frac{\log p}{n}}+2C_4\epsilon_2\\ =&2\left[C_3(3\log p+\log n)+(3\log p+\log n)^2\right]\sqrt{\frac{\log p}{n}}+2C_4\left(C_1+\sqrt{\frac{2\log p}{n}}\right).\end{aligned}$$

PROOF OF (19)

Then, we study (19). We consider a map defined as $G(\boldsymbol{\Delta}_S):=-\mathbf{H}_{SS}^{-1}\left[\boldsymbol{\nabla}_S F(\boldsymbol{\Theta}^*+\boldsymbol{\Delta}_S)+\lambda\hat{\mathbf{Z}}_S\right]+\boldsymbol{\Delta}_S$. If $\|\boldsymbol{\Delta}\|_\infty\leq r$, by Taylor expansion of $\boldsymbol{\nabla}_S F(\boldsymbol{\Theta}^*+\boldsymbol{\Delta})$ centered at $\boldsymbol{\nabla}_S F(\boldsymbol{\Theta}^*)$,

$$\begin{aligned}\|G(\boldsymbol{\Delta}_S)\|_\infty=&\left\|-\mathbf{H}_{SS}^{-1}\left[\boldsymbol{\nabla}_S F(\boldsymbol{\Theta}^*)+\mathbf{H}_{SS}\boldsymbol{\Delta}_S+\boldsymbol{R}_S(\boldsymbol{\Delta})+\lambda\hat{\mathbf{Z}}_S\right]+\boldsymbol{\Delta}_S\right\|_\infty=\left\|-\mathbf{H}_{SS}^{-1}\left(\boldsymbol{\nabla}_S F(\boldsymbol{\Theta}^*)+\boldsymbol{R}_S(\boldsymbol{\Delta})+\lambda\hat{\mathbf{Z}}_S\right)\right\|_\infty\\ \leq&\left\|\mathbf{H}_{SS}^{-1}\right\|_\infty\left(\|\boldsymbol{\nabla}_S F(\boldsymbol{\Theta}^*)\|_\infty+\|R_S(\boldsymbol{\Delta})\|_\infty+\lambda\|\hat{\mathbf{Z}}_S\|_\infty\right)\leq(C_5(\lambda+C_6r^2+\lambda)=C_5C_6r^2+2C_5\lambda,\end{aligned}$$

where the inequality is due to $\|\boldsymbol{\nabla}_S F(\boldsymbol{\Theta}^*)\|_\infty\leq\lambda$ conditioning on $E_1^c\cap E_2^c$ and according to (18). Then, based on the definition of r, we can derive the upper bound of $\|G(\boldsymbol{\Delta}_S)\|_\infty$ as $\|G(\boldsymbol{\Delta}_S)\|_\infty\leq r/2+r/2=r$.

Therefore, according to the fixed point theorem (Ortega and Rheinboldt, 2000; Yang and Ravikumar, 2011), there exists $\boldsymbol{\Delta}_S$ satisfying $G(\boldsymbol{\Delta}_S)=\boldsymbol{\Delta}_S$, which indicates $\boldsymbol{\nabla}_S F(\boldsymbol{\Theta}^*+\boldsymbol{\Delta})+\lambda\hat{\mathbf{Z}}_S=\mathbf{0}$. Considering that the optimal solution to (25) is unique, $\tilde{\boldsymbol{\Delta}}_S=\boldsymbol{\Delta}_S$, whose infinite norm is bounded by $\|\tilde{\boldsymbol{\Delta}}_S\|_\infty\leq r$, with probability larger than $1-\left((\exp\left(C_1+C_2/2\right)+8\right)p^{-2}+p^{-1/C_2}\big)$. □

Lemma 9. *Let $\hat{\boldsymbol{\Theta}}$ be an optimal solution to (12), and $\hat{\mathbf{Z}}$ be the corresponding dual solution. If $\hat{\mathbf{Z}}$ satisfies $\|\hat{\mathbf{Z}}_I\|_\infty<1$, then any given optimal solution to (12) $\tilde{\boldsymbol{\Theta}}$ satisfies $\tilde{\boldsymbol{\Theta}}_I=\mathbf{0}$. Moreover, if $\mathbf{H}_{SS}$ is positive definite, then the solution to (12) is unique.*

Proof. Specifically, following the same rationale as Lemma 1 in Wainwright 2009, Lemma 1 in Ravikumar et al. 2010, and Lemma 2 in Yang and Ravikumar 2011, we can derive Lemma 9 characterizing the optimal solution of (12). □

8.4. Proof of Theorem 1

The proof follows the primal-dual witness (PDW) technique, which is widely used in this line of research (Wainwright, 2009; Ravikumar et al., 2010; Yang and Ravikumar, 2011; Yang et al., 2015a). Specifically, by Lemma 9, we can prove the sparsistency by building an optimal solution to (12) satisfying $\|\hat{\mathbf{Z}}_I\|_\infty<1$, which is summarized as *strict dual feasibility* (SDF). To this end, we apply PDW to build a qualified optimal solution with the assumption that $\mathbf{H}_{SS}$ is positively definite.

SOLVE A RESTRICTED PROBLEM

First of all, we derive the KKT condition of (12):

$$\boldsymbol{\nabla}F(\hat{\boldsymbol{\Theta}})+\lambda\hat{\mathbf{Z}}=\mathbf{0}. \tag{24}$$

To construct an optimal optimal primal-dual pair solution, we define $\tilde{\boldsymbol{\Theta}}$ as an optimal solution to the restricted problem:

$$\tilde{\boldsymbol{\Theta}}:=\min_{\boldsymbol{\Theta}}F(\boldsymbol{\Theta})+\lambda\|\boldsymbol{\Theta}\|_1, \tag{25}$$

with $\boldsymbol{\Theta}_I=\mathbf{0}$, where $\tilde{\boldsymbol{\Theta}}$ is unique according to Lemma 9 with the assumption that $\mathbf{H}_{SS}\succ\mathbf{0}$. Denote the subgradient corresponding to $\tilde{\boldsymbol{\Theta}}$ as $\tilde{\mathbf{Z}}$. Then $(\tilde{\boldsymbol{\Theta}},\tilde{\mathbf{Z}})$ is optimal for the restricted problem (25). Therefore, $\tilde{\mathbf{Z}}_S$ can be determined according to the values of $\tilde{\boldsymbol{\Theta}}_S$ via the KKT conditions of (25). As a result,

$$\boldsymbol{\nabla}_S F(\tilde{\boldsymbol{\Theta}})+\lambda\tilde{\mathbf{Z}}_S=\mathbf{0}, \tag{26}$$

where $\boldsymbol{\nabla}_S$ represents the gradient components with respect to S. Furthermore, by letting $\hat{\boldsymbol{\Theta}}=\tilde{\boldsymbol{\Theta}}$, we determine $\tilde{\mathbf{Z}}_I$ according to (24). It remains to show that $\tilde{\mathbf{Z}}_I$ satisfies SDF.

CHECK SDF

Now, we demonstrate that $\tilde{\mathbf{\Theta}}$ and $\tilde{\mathbf{Z}}$ satisfy SDF. By (26), and by the Taylor expansion of $\nabla_S F(\tilde{\mathbf{\Theta}})$, we have that

$$\mathbf{H}_{SS}\tilde{\mathbf{\Delta}}_S + \nabla_S F(\mathbf{\Theta}^*) + \boldsymbol{R}_S(\tilde{\mathbf{\Delta}}) + \lambda\tilde{\mathbf{Z}}_S = \mathbf{0} \Rightarrow \tilde{\mathbf{\Delta}}_S = \mathbf{H}_{SS}^{-1}\left[-\nabla_S F(\mathbf{\Theta}^*) - \boldsymbol{R}_S(\tilde{\mathbf{\Delta}}) - \lambda\tilde{\mathbf{Z}}_S\right], \tag{27}$$

where $\tilde{\mathbf{\Delta}} := \tilde{\mathbf{\Theta}} - \mathbf{\Theta}^*$, $\boldsymbol{R}_S(\tilde{\mathbf{\Delta}})$ represents the components of $\boldsymbol{R}(\mathbf{\Delta})$ corresponding to S, and we have used the fact that $\mathbf{H}_{SS}$ is positive definite and hence invertible. By the definition of $\tilde{\mathbf{\Theta}}$ and $\tilde{\mathbf{Z}}$,

$$\nabla F(\tilde{\mathbf{\Theta}}) + \lambda\tilde{\mathbf{Z}} = \mathbf{0} \Rightarrow \nabla F(\mathbf{\Theta}^*) + \mathbf{H}\tilde{\mathbf{\Delta}} + \boldsymbol{R}(\tilde{\mathbf{\Delta}}) + \lambda\tilde{\mathbf{Z}} = \mathbf{0} \Rightarrow \nabla_I F(\tilde{\mathbf{\Theta}}) + \mathbf{H}_{IS}\tilde{\mathbf{\Delta}}_S + \boldsymbol{R}_I(\tilde{\mathbf{\Delta}}) + \lambda\tilde{\mathbf{Z}}_I = \mathbf{0}, \tag{28}$$

where $\boldsymbol{R}_I(\tilde{\mathbf{\Delta}})$ represents the components of $\boldsymbol{R}(\mathbf{\Delta})$ corresponding to I, and we have used the fact that $\tilde{\mathbf{\Delta}}_I = \mathbf{0}$ because $\tilde{\mathbf{\Theta}}_I = \mathbf{\Theta}^* = \mathbf{0}$. As a result,

$$\begin{aligned}
\lambda\|\tilde{\mathbf{Z}}_I\|_\infty =& \|-\mathbf{H}_{IS}\tilde{\mathbf{\Delta}}_S - \nabla_I F(\mathbf{\Theta}^*) - \boldsymbol{R}_I(\tilde{\mathbf{\Delta}})\|_\infty \\
\leq& \left\|\mathbf{H}_{IS}\mathbf{H}_{SS}^{-1}\left[-\nabla_S F(\mathbf{\Theta}^*) - \boldsymbol{R}_S(\tilde{\mathbf{\Delta}}) - \lambda\tilde{\mathbf{Z}}_S\right]\right\|_\infty + \|\nabla_I F(\mathbf{\Theta}^*) + \boldsymbol{R}_I(\tilde{\mathbf{\Delta}})\|_\infty \\
\leq& \left\|\mathbf{H}_{IS}\mathbf{H}_{SS}^{-1}\right\|_\infty \left\|\nabla_S F(\mathbf{\Theta}^*) + \boldsymbol{R}_S(\tilde{\mathbf{\Delta}})\right\|_\infty + \left\|\mathbf{H}_{IS}\mathbf{H}_{SS}^{-1}\right\|_\infty \left\|\lambda\tilde{\mathbf{Z}}_S\right\|_\infty + \|\nabla_I F(\mathbf{\Theta}^*) + \boldsymbol{R}_I(\tilde{\mathbf{\Delta}})\|_\infty \\
\leq& (1-\alpha)\left(\|\nabla_S F(\mathbf{\Theta}^*)\|_\infty + \|\boldsymbol{R}_S(\tilde{\mathbf{\Delta}})\|_\infty\right) + (1-\alpha)\lambda + \left(\|\nabla_I F(\mathbf{\Theta}^*)\|_\infty + \|\boldsymbol{R}_I(\tilde{\mathbf{\Delta}})\|_\infty\right) \\
\leq& (2-\alpha)\left(\|\nabla F(\mathbf{\Theta}^*)\|_\infty + \|\boldsymbol{R}(\tilde{\mathbf{\Delta}})\|_\infty\right) + (1-\alpha)\lambda,
\end{aligned} \tag{29}$$

where we have used (27) in the first inequality, and the third inequality is due to Assumption 3.

With (29), it remains to control $\|\nabla F(\mathbf{\Theta}^*)\|_\infty$ and $\|\boldsymbol{R}(\tilde{\mathbf{\Delta}})\|_\infty$. On one hand, according to Lemma 8 and the assumption on λ in Theorem 1, $\|\nabla F(\mathbf{\Theta}^*)\|_\infty \leq 2\left[3C_3\log p + C_3\log n + (3\log p + \log n)^2\right]\sqrt{\frac{\log p}{n}} + 2C_4\left(C_1 + \sqrt{\frac{2\log p}{n}}\right) \leq \frac{\alpha\lambda}{4}$, with probability larger than $1 - \left((\exp(C_1 + C_2/2) + 8)\, p^{-2} + p^{-1/C_2}\right)$.

On the other hand, according to Assumption 4 and Lemma 8,

$$\|\boldsymbol{R}(\tilde{\mathbf{\Delta}})\|_\infty \leq C_6\|\mathbf{\Delta}\|_\infty^2 \leq C_6 r^2 \leq C_6(4C_5\lambda)^2 = \lambda\frac{64C_5^2C_6}{\alpha}\frac{\alpha\lambda}{4} \leq \left(C_7\sqrt{\frac{\log^5 p}{n}}\right)\frac{64C_5^2C_6}{\alpha}\frac{\alpha\lambda}{4}, \tag{30}$$

where in the last inequality we have used the assumption $\lambda \propto \sqrt{\frac{\log^5 p}{n}}$ in Theorem 1, and hence there exists C_7 satisfying $\lambda \leq C_7\sqrt{\frac{\log^5 p}{n}}$. Therefore, when we choose $n \geq \left(64C_7C_5^2C_6/\alpha\right)^2\log^5 p$ as assumed in Theorem 1, then from (30), we can conclude that $\|\boldsymbol{R}(\tilde{\mathbf{\Delta}})\|_\infty \leq \frac{\alpha\lambda}{4}$. As a result, $\lambda\|\hat{\mathbf{Z}}_I\|_\infty$ can be bounded by $\lambda\|\tilde{\mathbf{Z}}_I\|_\infty < \alpha\lambda/2 + \alpha\lambda/2 + (1-\alpha)\lambda = \lambda$. Combined with Lemma 9, we demonstrate that any optimal solution of (12) satisfies $\tilde{\mathbf{\Theta}}_I = \mathbf{0}$. Furthermore, (19) controls the difference between the optimal solution of (12) and the real parameter by $\|\tilde{\mathbf{\Delta}}_S\|_\infty \leq r$, by the fact that $r \leq \|\mathbf{\Theta}_S^*\|_\infty$ in Theorem 1, $\hat{\mathbf{\Theta}}_S$ shares the same sign with $\mathbf{\Theta}_S^*$.