6 Proof of Proposition 2

The proof follows the development in [7], where they consider the case with $k = 2$. Denote $Q(\mathbf{x})$ as

$$Q(\mathbf{x}) = \log(P(\mathbf{x})/P(\mathbf{0})),$$

for any $\mathbf{x} = (x_1, \ldots, x_p) \in \mathcal{X}^p$. Given any $\mathbf{x}$, also denote $\bar{\mathbf{x}}_s := (x_1, \ldots, x_{s-1}, 0, x_{s+1}, \ldots, x_p)$.

Now, consider the following general form for $Q(\mathbf{x})$:

$$Q(\mathbf{x}) = \sum_{\mathbf{t_1} \in \mathbf{V}} \mathbf{x_{t_1}} \mathbf{G_{t_1}}(\mathbf{x_{t_1}}) + \ldots + \sum_{\mathbf{t_1}, \ldots, \mathbf{t_k} \in \mathbf{V}} \mathbf{x_{t_1}} \ldots \mathbf{x_{t_k}} \mathbf{G_{t_1, \ldots, t_k}}(\mathbf{x_{t_1}}, \ldots, \mathbf{x_{t_k}}), \tag{14}$$

since the joint distribution has atmost factors of size k. It can then be seen that

$$\begin{aligned} \exp(Q(\mathbf{x}) - Q(\bar{\mathbf{x}}_s)) &= P(\mathbf{x})/P(\bar{\mathbf{x}}_s) \\ &= P(x_s | x_1, \ldots, x_{s-1}, x_{s+1}, \ldots, x_p)/P(0 | x_1, \ldots, x_{s-1}, x_{s+1}, \ldots, x_p), \end{aligned} \tag{15}$$

where the first equality follows from the definition of Q, and the second equality follows from some algebra (See for instance Section 2 in [7]). Now, consider simplifications of both sides of (15). Given the form of $Q(\mathbf{x})$ in (14), we have

$$\begin{aligned} &Q(\mathbf{x}) - Q(\bar{\mathbf{x}}_1) = \\ &x_1 \left(G_1(x_1) + \sum_{t=2}^{p} x_t G_{1t}(x_1, x_t) + \sum_{t_2, \ldots, t_k \in \{2, \ldots, p\}} x_{t_2} \ldots x_{t_k} G_{1, t_2, \ldots, t_k}(x_1, \ldots, x_{t_k}) \right). \end{aligned} \tag{16}$$

Also, given the exponential family form of the node-conditional distribution specified in the theorem,

$$\log \frac{P(x_i | x_1, \ldots, x_{s-1}, x_{s+1}, \ldots, x_p)}{P(0 | x_1, \ldots, x_{s-1}, x_{s+1}, \ldots, x_p)} = E(x_{V \setminus s})(B(x_s) - B(0)) + (C(x_s) - C(0)). \tag{17}$$

Setting $x_t = 0$ for all $t \neq s$ in (15), and using the expressions for the left and right hand sides in (16) and (17), we obtain,

$$x_s G_s(x_s) = E(\mathbf{0})(B(x_s) - B(0)) + (C(x_s) - C(0)). \tag{18}$$

Setting $x_r = 0$ for all $r \notin \{s, t\}$,

$$x_s G_s(x_s) + x_s x_t G_{st}(x_s, x_t) = E(0, \ldots, x_t, \ldots, 0)(B(x_s) - B(0)) + (C(x_s) - C(0)). \tag{19}$$

Similarly,

$$x_t G_t(x_t) + x_s x_t G_{st}(x_s, x_t) = E(0, \ldots, x_s, \ldots, 0)(B(x_t) - B(0)) + (C(x_t) - C(0)). \tag{20}$$

From the above three equations, we obtain:

$$x_s x_t G_{st}(x_s, x_t) = \theta_{st}(B(x_s) - B(0))(B(x_t) - B(0)).$$

More generally, by considering non-zero triplets, and setting $x_r = 0$ for all $r \notin \{s, t, u\}$, we obtain,

$$\begin{aligned} &x_s G_s(x_s) + x_s x_t G_{st}(x_s, x_t) + x_s x_u G_{su}(x_s, x_u) + x_s x_t x_u G_{stu}(x_s, x_t, x_u) = \\ &E(0, \ldots, x_t, \ldots, x_u, \ldots, 0)(B(x_s) - B(0)) + (C(x_s) - C(0)), \end{aligned} \tag{21}$$

so that by a similar reasoning we can obtain

$$x_s x_t x_u G_{stu}(x_s, x_t, x_u) = \theta_{stu}(B(x_s) - B(0))(B(x_t) - B(0))(B(x_u) - B(0)).$$

More generally, we can show that

$$x_{t_1} \ldots x_{t_k} G_{t_1, \ldots, t_k}(x_{t_1}, \ldots, x_{t_k}) = \theta_{t_1, \ldots, t_k}(B(x_{t_1}) - B(0)) \ldots (B(x_{t_k}) - B(0)).$$

Thus, the k-th order factors in the joint distribution as specified in (14) are tensor products of $(B(x_s) - B(0))$, thus proving the statement of the theorem.

7 Proof of Proposition 3

Proof. The set of sufficient statistics in (10) do not include node-wise terms $\{X_s\}$; suppose we consider the model with these terms added. Suppose we zero-pad the true parameter $\theta^* \in \mathbb{R}^{\binom{p}{2}}$ to include zero weights over these node-wise terms; the resulting parameter would lie in $\mathbb{R}^{\binom{p}{2}+p}$; we will overload notation and denote this zero-padded parameter as θ^*. Similarly, given any $v \in \mathbb{R}^p$, we can treat these as weights over the node-wise terms, and zero-pad it to $\bar{v} \in \mathbb{R}^{\binom{p}{2}+p}$. Suppose that $\|v\|_2 = 1$; a simple calculation then shows that

$$\log \mathbb{E}[\exp(\langle v, X\rangle)] = A(\theta^* + \bar{v}) - A(\theta^*).$$

By a Taylor Series expansion, we have for some $r \in [0, 1]$,

$$\begin{aligned} A(\theta^* + \bar{v}) - A(\theta^*) &= \nabla A(\theta^*) \cdot \bar{v} + \frac{1}{2}\bar{v}^T \nabla^2 A(\theta^* + r\bar{v})\bar{v} \\ &\leq \|\mu^*\|_2 \|\bar{v}\|_2 + \frac{1}{2}|||\nabla^2 A(\theta^* + r\bar{v})||| \|\bar{v}\|_2^2 \\ &\leq \kappa_m + \frac{1}{2}\kappa_h, \end{aligned}$$

where we use the bounds in Assumption 3: noting that $\|r\bar{v}\|_2 \leq 1$, Assumption 3 yields $|||\nabla^2 A(\theta^* + r\bar{v})||| \leq \kappa_h$.

Thus, by the standard Chernoff bounding technique, for a unit norm vector v,

$$P(\langle v, x\rangle > a') \leq \exp(-a' + \kappa_m + \frac{1}{2}\kappa_h).$$

This can be extended for a non-unit norm vector u with norm $\|u\|_2 \leq c'$,

$$P(\langle u, x\rangle > a) \leq \exp(-\frac{a}{c'} + \kappa_m + \frac{1}{2}\kappa_h).$$

Thus, the statement in the proposition follows by setting $a = \delta \log \eta$:

$$P(\langle u, x\rangle > \delta \log \eta) \leq \exp(-\frac{\delta}{c'} \log \eta + \kappa_m + \frac{1}{2}\kappa_h) \leq c\eta^{-\delta/c'}.$$

where $c = \exp(\kappa_m + \frac{1}{2}\kappa_h)$. □

8 Proof of Proposition 4

Proof. The proof follows along similar lines as that of Proposition 3. Suppose we zero-pad the edge-weights in the true parameter $\theta^* \in \mathbb{R}^{\binom{p}{2}}$ to include a zero weight for sufficient statistic X_s^2; we will overload notation and denote this zero-padded parameter in $\mathbb{R}^{\binom{p}{2}+1}$ as θ^*. Similarly, let $\bar{v} \in \mathbb{R}^{\binom{p}{2}+1}$ be the zero-padded parameter with its last coordinate equal to $t \in \mathbb{R}$, so that $\|v\|_2 = t$. We then have A simple calculation shows that

$$\log \mathbb{E}[\exp(tX_s^2)] = A(\theta^* + \bar{v}) - A(\theta^*).$$

By a Taylor Series expansion, we have for some $r \in [0, 1]$,

$$\begin{aligned} A(\theta^* + \bar{v}) - A(\theta^*) &= \nabla A(\theta^*) \cdot \bar{v} + \frac{1}{2}\bar{v}^T \nabla^2 A(\theta^* + r\bar{v})\bar{v} \\ &\leq \mathbb{E}[X_t^2]\|\bar{v}\|_2 + \frac{1}{2}|||\nabla^2 A(\theta^* + r\bar{v})||| \|\bar{v}\|_2^2 \\ &\leq \kappa_v t + \frac{1}{2}\kappa_h t^2, \end{aligned}$$

where we use the bounds in Assumption 3. Note that $\|r\bar{v}\|_2 \leq t \leq 1$, Assumption 3 yields $|||\nabla^2 A(\theta^* + r\bar{v})||| \leq \kappa_h$.

Thus, again by the standard Chernoff bounding technique, for $t \leq 1$,

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_s^{(i)}\right)^2 \geq \delta\right) \leq \exp(-n\delta t + n\kappa_v t + \frac{n}{2}\kappa_h t^2)$$
$$\leq \exp(-n\frac{(\delta-\kappa_v)^2}{2\kappa_h^2}) \leq \exp(-n\frac{\delta^2}{4\kappa_h^2}),$$

for $\delta \leq \min\{2\kappa_v/3, \kappa_h + \kappa_v\}$. □

9 Proof of Theorem 1

Following the development in [3], we use the *primal-dual witness* method to prove the successful graph structure recovery. From the sub-gradient optimality condition of convex program (13), we have

$$\nabla\ell(\widehat{\theta}; X_1^n) + \lambda_n \widehat{Z} = 0 \tag{22}$$

where each entry of sub-gradient vector $\widehat{Z}$ satisfies the following property: $\widehat{Z}_{st} = \text{sign}(\widehat{\theta}_{st})$ if $\widehat{\theta}_{st} \neq 0$, and $|\widehat{Z}_{st}| \leq 1$ other wise.

Note that, for the regime $p \gg n$, the convex program (13) is not necessarily strictly convex, as a result there may be multiple optimal solutions. However, the following lemma presented in [3], is the key motivation of the primal-dual witness method:

Lemma 1. *Suppose that there exits an primal optimal solution $\widehat{\theta}$ with associated dual optimal solution $\widehat{Z}$ s.t. $\|\widehat{Z}_{S^c}\| < 1$. Then, any optimal solution $\tilde{\theta}$ should have $\tilde{\theta}_{S^c} = 0$.*

Based on this lemma, we prove the statement by the following steps: (see [3] for details)

(a) We set $\widehat{\theta}_S$ s.t. $\widehat{\theta}_S = \arg\min_{(\theta_S,0)\in\mathbb{R}^{p-1}}\{\ell(\theta; X_1^n) + \lambda_n\|\theta\|_1\}$, and $\widehat{Z}_S = \text{sign}(\widehat{\theta}_S)$.

(b) We set $\widehat{\theta}_{S^c} = 0$.

(c) We get $\widehat{Z}_{S^c}$ to satisfy the condition (22) with $\widehat{\theta}$ and $\widehat{Z}_S$.

(d) We check the dual feasibility condition and the sign consistency condition with high probability.

From now, we are going to show that $\|\widehat{Z}_S\|_\infty < 1$ with high probability. By some algebra the sub-gradient optimality condition (22) can be represented as $\nabla^2\ell(\theta^*; X_1^n)(\widehat{\theta}-\theta^*) = -\lambda_n\widehat{Z} + W^n + R^n$, where $W^n := -\nabla\ell(\theta^*; X_1^n)$ is the sample score function (that we will show is small with high probability), and R^n is the remainder term by applying coordinate-wise mean value theorem; $R_j^n = [\nabla^2\ell(\theta^*; X_1^n) - \nabla^2\ell(\bar{\theta}^{(j)}; X_1^n)]_j^T(\widehat{\theta} - \theta^*)$. Note that $\bar{\theta}^{(j)}$ is some vector on the line between $\widehat{\theta}$ and θ^*, and $[\cdot]_j^T$ is j-th row of matrix.

Using the notation for the Fisher Information matrix, we then have

$$Q^*(\widehat{\theta} - \theta^*) = -\lambda_n\widehat{Z} + W^n + R^n.$$

We will need the following lemmas that respectively control various terms in the above expression: the score term W^n, the deviation $\widehat{\theta}_S - \theta_S^*$, and the remainder term R^n.

Lemma 2. *Suppose that we set λ_n to satisfy $\frac{8(2-\alpha)}{\alpha}\sqrt{\kappa_3}\sqrt{\frac{\log p}{n^{1-\kappa_2}}} \leq \lambda_n \leq 4n^{\kappa_2}\kappa_3\frac{2-\alpha}{\alpha}\|\theta^*\|_2$ for some constant $\kappa_3 \leq \min\{2\kappa_v/3, 2\kappa_h + \kappa_v\}$. Suppose also that $n \geq \frac{8\kappa_h^2}{\kappa_3^2}\log p$. Then, for the mutual incoherence parameter $\alpha \in (0, 1]$,*

$$P\left(\frac{2-\alpha}{\lambda_n}\|W^n\|_\infty \leq \frac{\alpha}{4}\right) \geq 1 - \exp(-c_1 n) - c_2 p'^{-5/4} - \exp(-c_3 n)$$

where $p' := \max\{n, p\}$.

Lemma 3. *Suppose that* $\lambda_n d \leq \frac{\lambda_{\min}^2}{40\lambda_{\max} n^{\kappa_2} \log p'}$ *and* $\|W^n\|_\infty \leq \frac{\lambda_n}{4}$. *Then, we have*

$$P\left(\|\widehat{\theta}_S - \theta^*_S\|_2 \leq \frac{5}{\lambda_{\min}}\sqrt{d}\lambda_n\right) \geq 1 - c_1 p'^{-5/4} - c_2 n^{-2}. \tag{23}$$

for some constants $c_1, c_2 > 0$.

Lemma 4. *If* $\lambda_n d \leq \frac{\lambda_{\min}^2}{400\lambda_{\max} n^{\kappa_2} \log p'} \frac{\alpha}{2-\alpha}$, *and* $\|W^n\|_\infty \leq \frac{\lambda_n}{4}$, *then we have*

$$P\left(\frac{\|R^n\|_\infty}{\lambda_n} \leq \frac{\alpha}{4(2-\alpha)}\right) \geq 1 - c_1 p'^{-5/4} - c_2 p'^{-2}. \tag{24}$$

for some constants $c_1, c_2 > 0$.

The proof then follows from Lemmas 2-4 in a straightforward fashion, following [3]. Consider the choice of regularization parameter $\lambda_n = \frac{8(2-\alpha)}{\alpha}\sqrt{\kappa_3}\sqrt{\frac{\log p}{n^{1-\kappa_2}}}$. For a sample size greater than or equal to $\left(\frac{4\log p}{\kappa_3\|\theta^*\|_2^2}\right)^{\frac{1}{1+\kappa_2}}$, we satisfy the condition of Lemma 2, so that we may conclude, with high probability, $\|W^n\|_\infty \leq \frac{\lambda_n}{4}$. Moreover, for a sample size $n \geq L'\left[\left(\frac{2-\alpha}{\alpha}\right)^4 d^2(\log p')^3\right]^{1/(1-3\kappa_2)}$, and for some constant $L' > 0$, the conditions for Lemma 3 and 4 are satisfied and hence equations (24) and (23) holds with high probability.

Strict dual feasibility. Some algebra introduced in [3] yields that

$$\begin{aligned}
\|\widehat{Z}_{S^c}\|_\infty &\leq |||Q^*_{S^cS}(Q^*_{S^cS})^{-1}|||_\infty\Big[\frac{\|W^n_S\|_\infty}{\lambda_n} + \frac{\|R^n_S\|_\infty}{\lambda_n} + 1\Big] + \frac{\|W^n_S\|_\infty}{\lambda_n} + \frac{\|R^n_S\|_\infty}{\lambda_n} \\
&\leq (1-\alpha) + (2-\alpha)\Big[\frac{\|W^n\|_\infty}{\lambda_n} + \frac{\|R^n\|_\infty}{\lambda_n}\Big] \leq (1-\alpha) + \frac{\alpha}{4} + \frac{\alpha}{4} = 1 - \frac{\alpha}{2} < 1.
\end{aligned}$$

Correct sign recovery. For the successful sign recovery, it suffices to show that $\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq \frac{\theta^*_{\min}}{2}$. From Lemma 3, we have $\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq \|\widehat{\theta}_S - \theta^*_S\|_2 \leq \frac{5}{\lambda_{\min}}\sqrt{d}\lambda_n \leq \frac{\theta^*_{\min}}{2}$ as long as $\theta^*_{\min} \geq \frac{10}{\lambda_{\min}}\sqrt{d}\lambda_n$. This completes the proof.

10 Proof of Lemma 2

Proof. For a fixed $t \in \{1, ..., p-1\}$, we define $V_t^{(i)}$ for notational convenience so that

$$W_t^n = \frac{1}{n}\sum_{i=1}^n X_s^{(i)} X_t^{(i)} - X_t^{(i)} D'(\langle \theta^*, X_{\backslash s}^{(i)} \rangle) = \frac{1}{n}\sum_{i=1}^n V_t^{(i)}$$

Consider the upper bound on the moment generating function of $V_t^{(i)}$, conditioned on $X_{\backslash s}^{(i)}$,

$$\begin{aligned}
&\mathbb{E}[\exp(t' V_t^{(i)} | X_{\backslash s}^{(i)})] \\
&= \sum_{X_s^{(i)}} \exp\Big\{t'\Big[X_s^{(i)} X_t^{(i)} - X_t^{(i)} D'(\langle \theta^*, X_{\backslash s}^{(i)} \rangle)\Big] + (C(X_s^{(i)}) + X_s^{(i)} \langle \theta^*, X_{\backslash s}^{(i)} \rangle - D(\langle \theta^*, X_{\backslash s}^{(i)} \rangle))\Big\} \\
&= \exp\Big\{D(\langle \theta^*, X_{\backslash s}^{(i)} \rangle + t' X_t^{(i)}) - D(\langle \theta^*, X_{\backslash s}^{(i)} \rangle) - t' X_t^{(i)} D'(\langle \theta^*, X_{\backslash s}^{(i)} \rangle)\Big\} \\
&= \exp\Big\{\frac{t'^2}{2} {X_t^{(i)}}^2 D''(\langle \theta^*, X_{\backslash s}^{(i)} \rangle + v_i t' X_t^{(i)})\Big\} \quad \text{for some } v_i \in [0,1]
\end{aligned}$$

where the last equality holds by the second-order Taylor series expansion. Consequently, we have

$$\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}[\exp(t' V_t^{(i)}) | X_{\backslash s}^{(i)}] \leq \frac{1}{n}\sum_{i=1}^n \frac{t'^2}{2}\big(X_t^{(i)}\big)^2 D''(\langle \theta^*, X_{\backslash s}^{(i)} \rangle + v_i t' X_t^{(i)}).$$

Now, we define the event:

$$\xi_1 := \{\max_i |\langle \theta^*, X^{(i)}_{\backslash s}\rangle + v_i t' X^{(i)}_t| \le \kappa_1 \log p'\} \tag{25}$$

where κ_1 is the exponential family dependent constant in Assumption 5. Then, noting that $\langle \theta^*, X^{(i)}_{\backslash s}\rangle + v_i t' X^{(i)}_t$ is of the form $\langle u, X\rangle$ where $\|u\|_2 \le 2\|\theta^*\|_2$, provided that $t' \le \|\theta^*\|_2$,

$$P[\xi_1^c] \le c_2 n\, p'^{-\kappa_1/(2\|\theta^*\|_2)} \le c_2 p'^{-5/4}, \tag{26}$$

for some constant $c_2 > 0$, from the Proposition 3 and the union bound. By combining (26) with the Assumption 5, we obtain

$$\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}[\exp(t' V_t^{(i)})|X^{(i)}_{\backslash s}] \le \frac{n^{\kappa_2} t'^2}{2}\frac{1}{n}\sum_{i=1}^n \left(X_t^{(i)}\right)^2 \quad \text{for } t' \le \|\theta^*\|_2$$

with probability at least $1 - c_2 p'^{-5/4}$.

For each index t, the variables $\frac{1}{n}\left\{\left(X_t^{(i)}\right)^2\right\}_{i=1}^n$ satisfy the tail bound in Proposition 4. Let us define the event $\xi_2 := \left\{\max_{t=1,\ldots,p-1} \frac{1}{n}\sum_{i=1}^n \left(X_t^{(i)}\right)^2 \le \kappa_3\right\}$ for some constant $\kappa_3 \le \min\{2\kappa_v/3, 2\kappa_h + \kappa_v\}$. Then, we can establish the upper bound of probability $P[\xi_2^c]$ by a union bound,

$$P[\xi_2^c] \le 2\exp(-\frac{\kappa_3^2}{4\kappa_h^2} n + \log p) \le \exp(-c_3 n)$$

as long as $n \ge \frac{8\kappa_h^2}{\kappa_3^2}\log p$. Therefore, conditioned on ξ_1, ξ_2, the moment generating function is bounded as follows:

$$\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}[\exp(t' V_t^{(i)})|X^{(i)}_{\backslash s}, \xi_1, \xi_2] \le \frac{n^{\kappa_2}\kappa_3\, t'^2}{2} \quad \text{for } t' \le \|\theta^*\|_2.$$

The standard Chernoff bound technique implies that for any $\delta > 0$,

$$P\Big[\frac{1}{n}\sum_{i=1}^n |V_t^{(i)}| > \delta \mid \xi_1, \xi_2\Big] \le 2\exp\left(n\Big(\frac{n^{\kappa_2}\kappa_3\, t'^2}{2} - t'\delta\Big)\right) \quad \text{for } t' \le \|\theta^*\|_2.$$

Setting $t' = \frac{\delta}{n^{\kappa_2}\kappa_3}$:

$$P\Big[\frac{1}{n}\sum_{i=1}^n |V_t^{(i)}| > \delta \mid \xi_1, \xi_2\Big] \le 2\exp\left(-\frac{n\delta^2}{2n^{\kappa_2}\kappa_3}\right) \quad \text{for } \delta \le n^{\kappa_2}\kappa_3\|\theta^*\|_2.$$

Given the setting of the regularization parameter λ_n, we have $\frac{\alpha}{2-\alpha}\frac{\lambda_n}{4} \le n^{\kappa_2}\kappa_3\|\theta^*\|_2$ for n large enough; thus setting $\delta = \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}$:

$$P\Big[\frac{1}{n}\sum_{i=1}^n |V_t^{(i)}| > \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4} \mid \xi_1, \xi_2\Big] \le 2\exp\left(-\frac{\alpha^2}{(2-\alpha)^2}\frac{n\lambda_n^2}{32 n^{\kappa_2}\kappa_3}\right),$$

and by a union bound, we obtain

$$P\Big[\|W^n\|_\infty > \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4} \mid \xi_1, \xi_2\Big] \le 2\exp\left(-\frac{\alpha^2}{(2-\alpha)^2}\frac{n\lambda_n^2}{32 n^{\kappa_2}\kappa_3} + \log p\right).$$

Finally, provided that $\lambda_n \ge \frac{8(2-\alpha)}{\alpha}\sqrt{\kappa_3}\sqrt{\frac{\log p}{n^{1-\kappa_2}}}$, we obtain

$$P\Big[\|W^n\|_\infty > \frac{\alpha}{2-\alpha}\frac{\lambda_n}{4}\Big] \le \exp(-c_1 n) + c_2 p'^{-5/4} + \exp(-c_3 n),$$

where we use the fact that $P(X) \le P(X|\xi_1, \xi_2) + P(\xi_1^c) + P(\xi_2^c)$. □

11 Proof of Lemma 3

Proof. In order to establish the error bound $\|\widehat{\theta}_S - \theta^*_S\|_2 \leq B$ for some radius B, several works (e.g. [12, 3]) proved that it suffices to show $F(u_S) > 0$ for all $u_S := \widehat{\theta}_S - \theta^*_S$ s.t. $\|u_S\|_2 = B$ where

$$F(u_S) := \ell(\theta^*_S + u_S; X_1^n) - \ell(\theta^*_S; X_1^n) + \lambda_n(\|\theta^*_S + u_S\|_2 - \|\theta^*_S\|_2).$$

Note that for $\hat{u}_S := \widehat{\theta}_S - \theta^*_S$, $F(\hat{u}_S) = 0$. From now on, we show that $F(u_s)$ is strictly positive on the boundary of the ball with radius $B = M\lambda_n\sqrt{d}$ where $M > 0$ is a parameter that we will choose later in this proof. Some algebra yields

$$F(u_s) \geq (\lambda_n\sqrt{d})^2\Big\{ -\frac{1}{4}M + q^*M^2 - M\Big\} \tag{27}$$

where q^* is the minimum eigenvalue of $\nabla^2\ell(\theta^*_S + vu_S; X_1^n)$ for some $v \in [0,1]$. Moreover,

$$\begin{aligned}
q^* &:= \Lambda_{\min}\big(\nabla^2\ell(\theta^*_S + vu_S)\big) \\
&\geq \min_{v\in[0,1]} \Lambda_{\min}\big(\nabla^2\ell(\theta^*_S + vu_S)\big) \\
&\geq \Lambda_{\min}\big[\frac{1}{n}\sum_{i=1}^n D''(\langle\theta^*_S, X_S^{(i)}\rangle)X_S^{(i)}\big(X_S^{(i)}\big)^T\big] \\
&\qquad - \max_{v\in[0,1]} \Big\|\frac{1}{n}\sum_{i=1}^n D'''(\langle\theta^*_S + vu_S, X_S^{(i)}\rangle)\big(u_S^TX_S^{(i)}\big)X_S^{(i)}\big(X_S^{(i)}\big)^T\Big\|_2 \\
&\geq \lambda_{\min} - \max_{v\in[0,1]}\max_y \frac{1}{n}\sum_{i=1}^n |D'''(\langle\theta^*_S + vu_S, X_S^{(i)}\rangle)|\,|\langle u_S, X_S^{(i)}\rangle|\,\big(\langle X_S^{(i)}, y\rangle\big)^2
\end{aligned}$$

where $y \in \mathbb{R}^d$ s.t $\|y\|_2 = 1$.

Now define the events

$$\xi_3 := \{\max_{i=1,\ldots,n} |\langle\theta^*_S + vu_S, X_S^{(i)}\rangle| \leq \kappa_1\}, \quad \text{and} \tag{28}$$

$$\xi_4 := \{\max_{i,s} |X_s^{(i)}| \leq 4\log p'\}. \tag{29}$$

Given the setting of $B \leq \|\theta^*\|_2$, we have $P\big[\varepsilon_3^c\big] \leq c_1\, p'^{-5/4}$, and again, for all sample i, $D'''(\langle\theta^*_S + vu_S, X_S^{(i)}\rangle) \leq n^{\kappa_2}$ with high probability, similar as ξ_1 in the previous chapter. At the same time, by the Proposition 3, we obtain $P[\varepsilon_4^c] \leq c_2\, npp'^{-4} \leq c_2 p'^{-2}$. Note that since all the elements in vector $X_S^{(i)}$ is smaller than $4\log p'$, $|\langle u_S, X_S^{(i)}\rangle| \leq 4\log p'\sqrt{d}\|u_S\|_2 = 4\log p' M\lambda_n d$ for all i. Then, conditioned on ξ_3 and ξ_4,

$$q^* \geq \lambda_{\min} - 4\lambda_{\max}M\lambda_n d n^{\kappa_2}\log p',$$

As a result, assuming that $\lambda_n \leq \frac{\lambda_{\min}}{8\lambda_{\max}Mdn^{\kappa_2}\log p'}$, $q^* \geq \frac{\lambda_{\min}}{2}$. Finally, from (27), we obtain

$$F(u_s) \geq (\lambda_n\sqrt{d})^2\Big\{ -\frac{1}{4}M + \frac{\lambda_{\min}}{2}M^2 - M\Big\},$$

which is strictly positive for $M = \frac{5}{\lambda_{\min}}$.

Therefore, if $\lambda_n \leq \frac{\lambda_{\min}}{8\lambda_{\max}Mdn^{\kappa_2}\log p'} \leq \frac{\lambda^2_{\min}}{40\lambda_{\max}n^{\kappa_2}\log p'}$, then

$$\|\widehat{\theta}_S - \theta^*_S\|_2 \leq \frac{5}{\lambda_{\min}}\sqrt{d}\lambda_n,$$

which completes the proof. □

12 Proof of Lemma 4

Proof. In the proof, we are going to show that $\|R^n\|_\infty \leq 4n^{\kappa_2} \log p' \lambda_{\max} \|\widehat{\theta}_S - \theta^*_S\|_2^2$. Then, since the conditions of Lemma 4 are stronger than those of Lemma 3, from the result of Lemma 3, we can conclude that

$$\|R^n\|_\infty \leq \frac{100 n^{\kappa_2} \lambda_{\max} \log p'}{\lambda^2_{\min}} \lambda_n^2 d,$$

as claimed in Lemma 4.

From the definition of R^n, for a fixed $t \in \{1, ..., p-1\}$, R^n_t can be written as

$$\frac{1}{n} \sum_{i=1}^n \left[D''(\langle \theta^*, X^{(i)}_{\backslash s} \rangle) - D''(\langle \bar{\theta}^{(t)}, X^{(i)}_{\backslash s} \rangle) \right] \left[X^{(i)}_{\backslash s} \left(X^{(i)}_{\backslash s} \right)^T \right]_t^T [\widehat{\theta} - \theta^*]$$

where $\bar{\theta}^{(t)}$ is some point in the line between $\widehat{\theta}$ and θ^*, i.e., $\bar{\theta}^{(t)} = v_t \widehat{\theta} + (1 - v_t)\theta^*$ for $v_t \in [0, 1]$. By another application of the mean value theorem, we have

$$R^n_t = -\frac{1}{n} \sum_{i=1}^n \left\{ D'''(\langle \bar{\bar{\theta}}^{(t)}, X^{(i)}_{\backslash s} \rangle) X^{(i)}_t \right\} \left\{ v_t [\widehat{\theta} - \theta^*]^T X^{(i)}_{\backslash s} \left(X^{(i)}_{\backslash s} \right)^T [\widehat{\theta} - \theta^*] \right\}$$

for a some point $\bar{\bar{\theta}}^{(t)}$ between $\bar{\theta}^{(t)}$ and θ^*. Similarly in previous proofs, conditioned on the event ξ_3 and ξ_4 we obtain

$$|R^n_t| \leq \frac{4n^{\kappa_2} \log p'}{n} \sum_{i=1}^n \left\{ v_t [\widehat{\theta} - \theta^*]^T X^{(i)}_{\backslash s} \left(X^{(i)}_{\backslash s} \right)^T [\widehat{\theta} - \theta^*] \right\}.$$

Performing some algebra yields

$$|R^n_t| \leq 4n^{\kappa_2} \lambda_{\max} \log p' \, \|\widehat{\theta}_S - \theta^*_S\|_2^2, \quad \text{for all } t \in \{1, ..., p-1\}$$

with probability at least $1 - c_1 p'^{-5/4} - c_2 n^{-2}$ for some constants c_1 and c_2, which completes the proof. □

13 Data and Pre-processing for Genomic Network Examples

Our GLM graphical models were applied to two examples of non-Gaussian high-throughput genomic data to learn a Glioblastoma aberration network and a breast cancer meta-miRNA expression network. For the former, Level III array CGH Glioblastoma data [14] was downloaded from the Cancer Genomic Atlas (TCGA) portal (http://tcga-data.nci.nih.gov/tcga/). Array CGH data measures copy number variation, or the number of copies of a particular genomic region in a sample; there are normally two copies of a gene, maternal and paternal. The `Bioconductor` package `CNTools` [20] was used to convert the copy number information into a matrix structure with rows as overlapping genomic segments and columns as the subjects ($n = 461$). Each matrix element is categorized in one of three groups: amplified, normal, or deleted using defaults in `CNTools`. A sliding window algorithm was then applied across the genomic segments, merging segments in which the categories differed by less than 10% across all the subjects [21]. This resulted in a matrix of genomic regions by subjects. All genomic regions were ordered by their percentage of aberrations across all samples and the top 10% of these regions, 101 in total, were used in our analysis. Our processed data matrix was checked for batch effects by fitting a multinomial ANOVA model to each genomic region [22]; no significant batch effects were detected.

For the miRNA network, level III breast cancer miRNA expression [13] as measured by next generation sequencing was downloaded from the TCGA portal (http://tcga-data.nci.nih.gov/tcga/). MicroRNAs (miRNA) are short RNA fragments that are thought to be post-transcriptional regulators, enhancing or inhibiting gene expression. Measuring miRNA expression by high-throughput sequencing results in count data that is zero-inflated, highly skewed, and whose total count volume depends on experimental conditions [23]. Data was processed to be approximately Poisson by following the steps in [24]. In brief, the data was quantile corrected to adjust for sequencing depth [25],

the miRNAs with little variation across the samples, the bottom 50%, were filtered out, and the data was adjusted for possible overdispersion using a power transform and a goodness of fit test [23, 24]. The resulting data matrix consisting of 544 subjects and 262 miRNAs was tested for batch effects by fitting a Poisson ANOVA model [22]; only 4% of miRNAs were found to be associated with batch labels, and thus no significant batch association was detected. As Poisson graphical models are restricted to capture negative conditional relationships, we study the inhibitory effect of miRNAs through a meta-miRNA network. Meta-miRNAs were formed by clustering the miRNAs into tightly positively correlated groups, 32 in total, by using hierarchical cluster with average linkage. The mediod, or median centroid of each cluster, was taken to by the driver miRNA and formed the nodes of our meta-miRNA network.