Supplementary Material for “Learning the Network Structure of Heterogeneous Data via Pairwise Exponential Markov Random Fields”

Youngsuk Park
Stanford University
youngsuk@stanford.edu

David Hallac
Stanford University
hallac@stanford.edu

Stephen Boyd
Stanford University
boyd@stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

1 Proof of Theorem 3.2

We derive a convex upperbound on log-partition $A(\boldsymbol{\theta})$. By the convexity,

$$A(\boldsymbol{\theta}) = \sup_{\boldsymbol{\mu}\in\mathcal{Y}} \langle \boldsymbol{\mu}, \boldsymbol{\theta}\rangle - A^*(\boldsymbol{\mu}) \tag{1}$$

where $A^*(\boldsymbol{\mu})$ is the conjugate function $A^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta}\in\mathcal{Y}}\langle \boldsymbol{\mu}, \boldsymbol{\theta}\rangle - A(\boldsymbol{\theta})$. One key fact is that $A^*(\boldsymbol{\mu})$ can be expressed as the Shannon entropy as follows [3]:

$$A^*(\mu) = \begin{cases} -H(p(x;\boldsymbol{\theta}(\boldsymbol{\mu}))) & \boldsymbol{\mu}\in\mathcal{M}^o(\boldsymbol{B}) \\ -\lim_{\{\boldsymbol{\mu}_s\}\to\boldsymbol{\mu}} H(p(x;\boldsymbol{\theta}(\boldsymbol{\mu}_s))) & \mu\in\mathbf{bd}(\mathcal{M}(\boldsymbol{B})) \\ +\infty & \text{otherwise} \end{cases}.$$

Here $\mathcal{M}(\boldsymbol{B})$, $\mathcal{M}^o$, and $\mathbf{bd}(\mathcal{M})$ is the *mean parameter* (the set of realizable expected sufficient statistic), its interior, is its boundary of $\mathcal{M}$ respectively. And $\boldsymbol{\theta}(\boldsymbol{\mu})$ is the unique natural parameter satisfying $\boldsymbol{\mu} = \sum_{x\in\mathcal{X}} p(x;\boldsymbol{\theta})\boldsymbol{B}(x)$ for $\boldsymbol{\mu}\in\mathcal{M}^o(\boldsymbol{B})$, and $\{\boldsymbol{\mu}_s\}\in\mathcal{M}^o(\boldsymbol{B})$ is a sequence converging to $\boldsymbol{\mu}\in\mathbf{bd}(\mathcal{M}(\boldsymbol{B}))$.

Recall that we denote $d = \sum_{r=1}^p m_r$ and $b_{node}(x) = \mathbf{vec}[B_1(X_1),\ldots,B_p(X_p)]\in\mathbf{R}^d$.

Express $H(X)$ **as** $H(b_{node}(X))$**.** First, we will derive an upperbound of $H(p(x;\theta)) \equiv H(X)$ in terms of $H(b_{node}(X))$. By the chain rule for entropy [2]

$$\begin{aligned} H(X, b_{node}(X)) &= H(X) + H(b_{node}(X) \mid X) \\ &\overset{(a)}{=} H(X) \end{aligned}$$

where (a) holds because the entropy is zero when the variable is deterministic on the condition, i.e., the function of the condition [2]. From the other direction of chain rule,

$$\begin{aligned} H(X, b_{node}(X)) &= H(b_{node}(X)) + H(b_{node}(X) \mid X) \\ &\overset{(b)}{=} H(b_{node}(X)) + \sum_{r=1}^p H(X_r \mid b_{node}(X), X_1, \ldots, X_{r-1}) \\ &\overset{(c)}{\leq} H(b_{node}(X)) + \sum_{r=1}^p H(X_r \mid B_r(X)) \end{aligned}$$

where (b) holds by applying chain rule successively, and (c) holds because conditioning reduces entropy. Note that, for a known exponential family distribution, $H(X_r \mid B_r(X))$ is constant. For examples in a Gaussian, Dirichlet, Gamma, Wishart, X_r is also a function of $B_r(X_r)$, meaning $H(X_r \mid B_r(X)) = 0$. For a Laplacian distribution, $H(X_r \mid B_r(X)) = H(\mathbf{Sign}(X_r)) = 1$. Therefore, we can conclude that $H(X) = H(b_{node}(X)) + C_0$ where $C_0 = \sum_{r=1}^p H(X_r \mid B_r(X))$ is the constant determined by the types of nodes.

Now we derive an upper bound on entropy $H(b_{node}(X))$. We will utilize the fact that, among any continuous random vectors with the same covariance matrix, Gaussian random vector maximized the entropy. To do so, we need to construct an additive and independent random vector $U\in\mathbf{R}^d$ so that $b_{node}(X)+U$ becomes a continuous random vector and its entropy $H(b_{node}(X)+U)$ is closely related to $H(b_{node}(X))$.

Construct U**.** For each discrete node $r\in\mathcal{I}_D\subseteq\{1,\ldots,p\}$, define the distance $c_r = \inf_{a\neq b\in\mathcal{X}_r} \|B_r(a)-B_r(b)\|_\infty > 0$ in the domain of its sufficient statistic $B_r(\mathcal{X}_r)$. we define $c_r = 0$ otherwise.

Now, we construct a d-dimensional random vector $U = \mathbf{vec}[U_1,\ldots,U_p]$ where each element of $U_r = [U_{r1},\ldots,U_{rm_r}]^T\in\mathbf{R}^{m_r}$ is independently distributed as

$$U_{ri_r} \sim \begin{cases} \text{unif}[-c_r/2, c_r/2] & r\in\mathcal{I}_D \text{ and } i_r\in\{1,\ldots,m_r\} \\ 0 & \text{otherwise} \end{cases},$$

and is independent to $b_{node}(X)$. Since U has a sufficiently narrow range and is independent on $b_{node}(X)$, $b_{node}(X)$ is uniquely determined by the (continuous) random vector $b_{node}(X)+U$.

Express $H(b_{node}(X))$ **as** $H(b_{node}(X)+U)$**.** By the chain rule,

$$H(b_{node}(X)+U, b_{node}(X)) \overset{(d)}{=} H(b_{node}(X)+U),$$

where (d) holds because $b_{node}(X)$ is deterministic under $b_{node}(X)+U$. On the other hand,

$$\begin{aligned} &H(b_{node}(X)+U, b_{node}(X)) \\ &\qquad = H(b_{node}(X)) + H(b_{node}(X)+U \mid b_{node}(X)) \\ &\qquad \overset{(e)}{=} H(b_{node}(X)) + H(U) \\ &\qquad = H(b_{node}(X)) + \sum_{r\in\mathcal{I}_D}\big(m_r\log c_r\big), \end{aligned}$$

where (e) holds because Shannon Entropy is invariant under a transition and under a condition on independent random vectors.

Therefore we conclude $H(b_{node}(X)) = H(b_{node}(X)+U) - \sum_{r\in\mathcal{I}_D} m_r \log c$.

Entropy Bound on $H(b_{node}(X)+U)$. Since (differential) Shannon entropy of any continuous random vector is upper bounded by that of a Gaussian random vector with the same covariance matrix,

$$
\begin{aligned}
&H(b_{node}(X)+U) \leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det \mathbf{Cov}[b_{node}(X)+U] \\
&\overset{(f)}{=} \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det \Bigg(\\
&\begin{bmatrix} 1 & \mathbf{E}[b_{node}(X)+U] \\ \mathbf{E}[b_{node}(X)+U]^T & \mathbf{E}[(b_{node}(X)+U)(b_{node}(X)+U)^T] \end{bmatrix} \Bigg) \\
&\overset{(g)}{=} \frac{1}{2}\log\det \Bigg(\begin{bmatrix} 1 & \mathbf{E}[b_{node}(X)] \\ \mathbf{E}[b_{node}(X)]^T & \mathbf{E}[b_{node}(X) b_{node}(X)]^T \end{bmatrix} \\
&\qquad + \mathbf{diag}\left([0, l_1, \ldots, l_p]\right) \Bigg) + \frac{d}{2}\log(2\pi e),
\end{aligned}
$$

where (f) holds by the Shur complement [1], and (g) holds because U is independent on b_{node} with statistics $\mathbf{E}[U] = 0$ and $\mathbf{E}[U_r U_r^T] = \mathbf{diag}(l_r)$ with $l_r = \frac{c_r^3}{12}\mathbf{1}_{m_r}$. Here, $\mathbf{1}_{m_r}$ is a 1-valued vector in $\mathbf{R}^{m_r}$.

Note that each $\boldsymbol{\mu} \in \mathcal{M}(B)$ equals to $\mathbf{E}[\boldsymbol{B}(X)]$ under some valid $p(\cdot)$, and $\boldsymbol{\mu}$ is composed of $\mathbf{E}[\{b_{node}(X), b_{node}(X)b_{node}(X)^T\}]$. By using a map M_ν we defined, it can simply be expressed as

$$
\begin{bmatrix} 1 & \mathbf{E}[b_{node}(X)] \\ \mathbf{E}[b_{node}(X)]^T & \mathbf{E}[b_{node}(X) b_{node}(X)]^T \end{bmatrix} = M_1[\boldsymbol{\mu}].
$$

Finally,

$$
\begin{aligned}
A(\boldsymbol{\theta}) &= \max_{\boldsymbol{\mu}\in\mathcal{M}(\boldsymbol{B})} \{\langle \boldsymbol{\mu}, \boldsymbol{\theta}\rangle - A^*(\boldsymbol{\mu})\} \\
&\leq \max_{\boldsymbol{\mu}\in\mathcal{M}(\boldsymbol{B})} \left\{ \langle \boldsymbol{\mu}, \boldsymbol{\theta}\rangle + \frac{1}{2}\log\det\left(M_1[\boldsymbol{\mu}] + D\right) \right\} + f_1.
\end{aligned}
$$

where $D = \mathbf{diag}\left([0, l_1, \ldots, l_p]\right)$ and a constant $f_1 = \frac{d}{2}\log(2\pi e) - \sum_{r\in\mathcal{I}_D} m_r \log c_r + C_0$ determined by the types and dimension of nodes.

2 Proof of Corollary 3.2

By taking the relaxation of the dual, we can convert the high-dimensional problem from Theorem 3.1 into the following tractable form. Here, we introduce $\boldsymbol{\theta}' = \{\frac{\theta_1}{2}, \ldots, \frac{\theta_1}{2}, \Theta_{11}, \ldots, \Theta_{pp}\}$, a slight variant of $\boldsymbol{\theta}$, for an algebraic simplicity on the derivation.

$$
\begin{aligned}
A(\theta) &\leq \max_{\boldsymbol{\mu}\in\mathcal{M}(B)} \left\{ \langle \boldsymbol{\mu}, \boldsymbol{\theta}\rangle + \frac{1}{2}\log\det\left(M_1[\boldsymbol{\mu}] + D\right) + f_1 \right\} \\
&= \max_{\boldsymbol{\mu}\in\mathcal{M}(B)} \left\{ \langle M_1[\boldsymbol{\mu}], M_1[\boldsymbol{\theta}']\rangle + \frac{1}{2}\log\det\left(M_1[\boldsymbol{\mu}] + D\right) + f_1 - 1 \right\} \\
&\leq \frac{1}{2} \max_{X\succeq D, X_{11}=1} \left\{ \langle X - D, 2M_1[\boldsymbol{\theta}']\rangle + \log\det X \right\} + f_1 - 1 \\
&\leq \frac{1}{2} \max_{X\succeq 0, X_{11}=1} \left\{ \langle X, 2M_1[\boldsymbol{\theta}']\rangle + \log\det X \right\} - \big\langle D, M_1[\boldsymbol{\theta}]\big\rangle + f_1 - 1 \\
&= \frac{1}{2} \max_{X\succeq 0} \min_{v\in\mathbf{R}} \left\{ \langle X, 2M_1[\boldsymbol{\theta}']\rangle + \frac{1}{2}\log\det X + v(X_{11} - 1) \right\} \\
&\qquad\qquad - \big\langle D, M_1[\boldsymbol{\theta}']\big\rangle + f_1 - 1 \\
&= \frac{1}{2} \min_{v\in\mathbf{R}} \left(\max_{X\succeq 0} \left\{ \big\langle X, 2M_{1+\nu/2}[\boldsymbol{\theta}']\big\rangle + \log\det X \right\} - v \right) \\
&\qquad\qquad - \big\langle D, M_1[\boldsymbol{\theta}']\big\rangle + f_1 - \frac{3}{2} \\
&= \frac{1}{2} \min_{v\in\mathbf{R}} \left\{ -\log\det\left(-2M_{1+\nu/2}[\boldsymbol{\theta}']\right) - v \right\} \\
&\qquad\qquad - \big\langle D, M_1[\boldsymbol{\theta}']\big\rangle + f_2,
\end{aligned}
$$

where $f_2 = \frac{d}{2}\log(2\pi e) - \sum_{r\in\mathcal{I}_D} m_r \log c_r + C_0 - \frac{3}{2} + \frac{d}{2}$.

3 Proof of Theorem 3.4

By combining the upper bound $\tilde{A}(\theta)$ of Corollary 3.2 with the regularized maximum likelihood equation in the paper, problem (4), we find that $\min_{\boldsymbol{\theta}} -\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\theta}\rangle + A(\boldsymbol{\theta}) + R_\lambda(\boldsymbol{\theta})$ is upper bounded by the following:

$$
\begin{aligned}
&\min_{\boldsymbol{\theta}} -\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\theta}\rangle + A(\boldsymbol{\theta}) + R_\lambda(\boldsymbol{\theta}) \\
&= \min_{\boldsymbol{\theta}\in\mathcal{Y}} \Bigg(-\langle M_1[\hat{\boldsymbol{\mu}}], M_1[\boldsymbol{\theta}']\rangle + \frac{1}{2}\min_{\nu\in\mathbf{R}} \Bigg\{ -\log\det\left(-2M_{1+\nu/2}[\boldsymbol{\theta}']\right) \\
&\qquad\qquad - \nu \Bigg\} - \big\langle D, M_1[\boldsymbol{\theta}']\big\rangle + R_\lambda(\boldsymbol{\theta}') \Bigg) + f_2 + 1 \\
&= \frac{1}{2} \min_{\nu\in\mathbf{R}, \boldsymbol{\theta}\in\mathcal{Y}} \Bigg\{ -\langle M_1[\hat{\boldsymbol{\mu}}] + D, 2M_{1+\nu/2}[\boldsymbol{\theta}']\rangle \\
&\qquad\qquad - \log\det\left(-2M_{1+\nu/2}[\boldsymbol{\theta}']\right) + 2R_\lambda(\boldsymbol{\theta}) \Bigg\} + f_2 + 1 \\
&= \frac{1}{2} \min_{\boldsymbol{\Theta}\in S^{d+1}} \left\{ \big\langle M_1[\hat{\boldsymbol{\mu}}] + D, \boldsymbol{\Theta}\big\rangle - \log\det\boldsymbol{\Theta} + R_{2\lambda}(\boldsymbol{\Theta}) \right\} + f_3
\end{aligned}
$$

where we replaced $-2M_{1+\nu/2}[\boldsymbol{\theta}']$ with $\boldsymbol{\Theta}$ and $R_\lambda(\boldsymbol{\theta})$ with $R_\lambda(\boldsymbol{\Theta})$.

4 Proof of Lemma 4.1

For notational simplicity, we denote $\mathbf{vec}[A] = \bar{A}$ as a vector by stacking all the columns in consistent order. We define $B_{st:ij} = [B_s(X_s)B_t(X_t)]_{ij}$ for $1 \le s,t \le p$, $1 \le i \le m_s$, and $1 \le j \le m_t$. Likewise, denote $B_{r:k} = [B_r(X_r)]_k$ for $1 \le r \le p$, and $1 \le k \le m_r$. The same notation on any elements in $\mathcal{Y}$ is applied. We define a term $W^k = M_1[\boldsymbol{B}(x^k)] - M_1[\mathbf{E}[\boldsymbol{B}(X)]]$ and its average $\hat{W} = M_1[\hat{\boldsymbol{\mu}}] - M_1[\mathbf{E}[\boldsymbol{B}(X)]]$ for n-samples $\mathbf{x} = \{x^1, \ldots, x^n\}$. Note that $\{W_k\}$'s are the i.i.d.

We use the Chernoff bound on $\hat{W}$. By abusing some notation, $\boldsymbol{\theta} + E_{st:ij}$ is denoted as the unit increment of the $\theta_{st:ij}$ element.

$$\begin{aligned}\mathbf{Pr}\left[\left\|\hat{W}\right\|_\infty > \delta_n\right] &\le 2\exp\left[(-nt\delta_n) + \sum_{k=1}^{n} \log\left(\mathbf{M}_{W^k}(t)\right)\right] \\ &\le 2\exp\left[-n\left(t\delta_n - \log \mathbf{M}_{W^1}(t)\right)\right],\end{aligned}$$

where $\mathbf{M}_W(t)$ is denoted as the moment generating function of W. We will get the upper bound on $\mathbf{M}_{W^1}(t) = \mathbf{E}[\exp(M_1[\boldsymbol{B}(X)])]$ by following

$$\begin{aligned}\mathbf{E}[\exp(B_{st:ij})] &= \int \exp[tB_{st:ij} + \langle \boldsymbol{B}(x), \boldsymbol{\theta}\rangle + C(x)\nu(dx) \\ &\qquad - A(\boldsymbol{\theta}) - t\,\mathbf{E}[B_{st:ij}]]\nu(dx) \\ &= \int \exp[\langle B(x), \boldsymbol{\theta} + tE_{st:ij}\rangle + C(x) - A(\boldsymbol{\theta} + tE_{st:ij})\nu(dx) \\ &\qquad + A(\boldsymbol{\theta} + tE_{st:ij}) - A(\boldsymbol{\theta}) - t\,\mathbf{E}[B_{st:ij}]]\nu(dx) \\ &= \int \exp[A(\boldsymbol{\theta} + tE_{st:ij}) - A(\boldsymbol{\theta}) - t\,\mathbf{E}[B_{st:ij}]]\nu(dx) \\ &\overset{(a)}{=} \int \exp[t\nabla\bar{A}(\boldsymbol{\theta})^T\bar{E}_{ab:ij} + \frac{1}{2}vt^2\bar{E}_{ab:ij}^T\nabla^2 A(\bar{\boldsymbol{\theta}})\bar{E}_{ab:ij} - t\,\mathbf{E}[B_{st:ij}]]\nu(dx) \\ &\overset{(b)}{=} \int \exp[t\,\mathbf{E}[\bar{\boldsymbol{B}}(X)]^T\bar{E}_{ab:ij} + \frac{1}{2}vt^2\bar{E}_{ab:ij}^T\,\mathbf{Cov}(\bar{\boldsymbol{B}}(X))\bar{E}_{ab:ij} \\ &\qquad - t\,\mathbf{E}[B_{st:ij}]]\nu(dx) \\ &\le \exp\left(\frac{1}{2}\kappa_{B_{\mathbf{Cov}[B]}}t^2\right),\end{aligned}$$

where (a) holds due to Taylor expansion at $\boldsymbol{\theta}$ with the value $0 \le v \le 1$, (b) holds due to properties of first and second derivative of the log-partition function for a exponential family, and (c) holds due to $\|E_{ab;ij}\|_2 = 1$.

Likewise $\mathbf{E}[\exp(B_{r:k})] \le \exp\left(\frac{1}{2}\kappa_{B_{\mathbf{Cov}[B]}}t^2\right)$.

Therefore, from the $\left\|\hat{W}\right\|_{\infty,2} \le m_{max}\left\|\hat{W}\right\|_\infty$, we get

$$\begin{aligned}\mathbf{Pr}\left[\|W\|_{\infty,2} > \delta_n\right] &\le m_{max}^2\,\mathbf{Pr}\left[\|W\|_\infty > \frac{\delta_n}{m_{max}}\right] \\ &\le m_{max}^2 p^2 2\exp[-n(t\frac{\delta_n}{m_{max}} - \frac{1}{2}\kappa_{\mathbf{Cov}[B]}t^2)].\end{aligned}$$

By setting $t = \delta_n/(m_{max}\kappa_{\mathbf{Cov}[B]})$, we get

$$\begin{aligned}&\mathbf{Pr}\left[\|W\|_{\infty,2} > \delta_n\right] \\ &\quad\le 2\exp\left[-\frac{n}{2m_{max}^2\kappa_{\mathbf{Cov}[B]}}\left(\delta_n^2 - \frac{2m_{max}^2\kappa_{\mathbf{Cov}[B]}\log(m_{max}p)}{n}\right)\right] \\ &\quad\le e^{-c_1 n},\end{aligned}$$

where the last inequality holds for for $\delta_n \ge 2m_{max}\sqrt{\frac{2\kappa_{\mathbf{Cov}[B]}\log(m_{max}p)}{n}}$ and for some positive universal constant c_1.

5 Proof of Theorem 4.2

We use the primal-dual witness approach developed by Ravikumar et al [19] and keep their notation. The main difference is that we analyze the optimality condition for the weighted group lasso penalty.

By the optimality condition for *group graphical lasso*, the estimator $\boldsymbol{\Theta}$ must satisfy

$$M_1[\hat{\boldsymbol{\mu}}] - \boldsymbol{\Theta} + \lambda_n\gamma_w \circ Z = 0 \qquad (2)$$

where $Z \in \mathbf{R}^{(d+1)\times(d+1)}$ is the subgradient at $\boldsymbol{\Theta}$ with

$$Z_{st:ij} = \begin{cases} 0 & \text{if} \quad \Theta_{st} = \mathbf{0}, \\ \frac{(\Theta_{st})_{ij}}{\|\Theta_{st}\|_F} & \text{otherwise} \end{cases}, \qquad (3)$$

for the block off-diagoanl parts and with a zero value for all other elements in the matrix. Note that this implies $\|Z\|_{\infty,2} \le 1$ and $\langle\Theta_{st}, Z_{st}\rangle_F \le \|\Theta_{st}\|_F$.

From the primal-dual witness approach, let S be the set of edges (excluding self-edges) and S be the set of non-edges.

Define the radius r that will eventually be used as an error measure $\|\Delta\|_{\infty,2} = \left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_{\infty,2}$.

5.1 Three Lemmas

In order to prove Theorem 4.2, we first present three Lemmas. In Lemma 4.1, the remainder of the second order Taylor expansion on $g(\boldsymbol{\Theta})|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^*}$ is expressed with the error $\|\Delta\|_{\infty,2}$. Then, we derive the sufficient condition of W, $R(\Delta)$ satisfying the optimality condition (1) in Lemma 4.2. Lastly, we derive the condition of the radius r which upper bounds the error $\|\Delta\|_{\infty,2}$.

Lemma 5.1 *Under the assumption* $r := \|\Delta\|_{\infty,2} < 1/(3\kappa_{\Gamma^*}d\sqrt{m_{max}})$*, we get*

$$\|R(\Delta)\|_{\infty,2} \le \frac{3}{2}d\,\|\Delta\|_{\infty,2}^2\,\kappa_{\Sigma^*}^3. \qquad (4)$$

Proof. For $\|u\|_{\infty,2} = 1$

$$\begin{aligned}|||\Delta|||_{\infty,2} &= \sup_{sin\{0,1,\ldots,p\}} \|\Delta_{s,:}u\|_2 \\ &= \sup_{s\in\{0,1,\ldots,p\}, i\in\{1,\ldots,m_s\}} \sqrt{m_{max}}\left\|\Delta_{(s,i),:}u\right\|_2 \\ &= \sup_{s\in\{0,1,\ldots,p\}, i\in\{1,\ldots,m_s\}} \sqrt{m_{max}}\sqrt{\sum_t \left\|\Delta_{(s,i),t}u_t\right\|_2^2} \\ &= \sup_{s\in\{0,1,\ldots,p\}, i\in\{1,\ldots,m_s\}} \sqrt{m_{max}}\sqrt{\sum_t \left\|\Delta_{(s,i),t}\right\|_2^2\|u_t\|_2^2} \\ &= \sup_{s\in\{0,1,\ldots,p\}, i\in\{1,\ldots,m_s\}} \sqrt{m_{max}}\sqrt{\sum_t \left\|\Delta_{(s,i),t}\right\|_2^2} \\ &= d\sqrt{m_{max}}\,\|\Delta\|_{\infty,2}\end{aligned}$$

For $J = \sum_{k=0}^{\infty}(-1)^k(\boldsymbol{\Theta}^{*-1}\Delta)^k$

$$\left|\left|\left|J^T\right|\right|\right|_{\infty,2} \le \sum_{k=0}^{\infty}(-1)^k\left|\left|\left|\boldsymbol{\Theta}^{*-1}\Delta\right|\right|\right|_{\infty,2}^k \le \frac{1}{1 - |||\boldsymbol{\Theta}^{*-1}|||_{\infty,2}|||\Delta|||_{\infty,2}} \le \frac{3}{2}$$

This claim is immediately applied to the Lemma 5 of Ravikumar et al.

$$(\mathbf{\Theta}^* + \Delta)^{-1} = \mathbf{\Theta}^{*-1} - \mathbf{\Theta}^{*-1}\Delta\mathbf{\Theta}^{*-1} + \mathbf{\Theta}^{*-1}\Delta\mathbf{\Theta}^{*-1}\Delta J\mathbf{\Theta}^{*-1}$$

$$R(\Delta) = (\mathbf{\Theta}^* + \Delta)^{-1} - \mathbf{\Theta}^{*-1} + \mathbf{\Theta}^{*-1}\Delta\mathbf{\Theta}^{*-1} = \mathbf{\Theta}^{*-1}\Delta\mathbf{\Theta}^{*-1}\Delta J\mathbf{\Theta}^{*-1}$$

$$\begin{aligned}|||R(\Delta)|||_{\infty,2} &\leq \left|\left|\left|\mathbf{\Theta}^{*-1}\right|\right|\right|_{\infty,2} |||\Delta|||_{\infty,2} \left|\left|\left|\mathbf{\Theta}^{*-1}\right|\right|\right|_{\infty,2} \left\|\mathbf{\Theta}^{*-1}J^T\Delta\right\|_{\infty,2} \\ &\leq \left|\left|\left|\mathbf{\Theta}^{*-1}\right|\right|\right|_{\infty,2}^2 |||\Delta|||_{\infty,2} \left|\left|\left|J^T\right|\right|\right|_{\infty,2} \|\Delta\|_{\infty,2} \\ &\leq \left|\left|\left|\mathbf{\Theta}^{*-1}\right|\right|\right|_{\infty,2}^3 \left|\left|\left|J^T\right|\right|\right|_{\infty,2} |||\Delta|||_{\infty,2} \|\Delta\|_{\infty,2}\end{aligned}$$

Lemma 5.2 *Suppose that the incoherence condition holds and* $\max\{\|W\|_{\infty,2}, \|R(\Delta)\|_{\infty,2}\} \leq \frac{w_{min}w_{max}\alpha}{4(w_{max}+w_{min})}\lambda_n$. *Then* $\|Z_{S^c}\|_{\infty,2} < 1$.

Proof. By utilizing the fact that $\Delta_{S^c} = \mathbf{0}$, we can derive

$$\Gamma^*_{SS}\bar{D}_S - \bar{R}_S + \bar{W}_S + \lambda_n\bar{\gamma}_w \circ \bar{Z}_S = 0 \quad (5)$$

$$\Gamma^*_{S^cS}\bar{D}_S - \bar{R}_{S^c} + \bar{W}_{S^c} + \lambda_n\bar{\gamma}_w \circ \bar{Z}_{S^c} = 0. \quad (6)$$

Combining two equation above gives

$$\begin{aligned}&\lambda_n\bar{\gamma}_w \otimes \bar{Z}_{S^c} \\ &= -\Gamma^*_{S^cS}\bar{D}_S + \bar{R}_{S^c} - \bar{W}_{S^c} \\ &= -\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}(\bar{W}_S - \bar{R}_S) + \lambda_n\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}(\bar{\gamma}_w \circ \bar{Z}_{S^c}) \\ &\quad + (\bar{R}_{S^c} - \bar{W}_{S^c}),\end{aligned}$$

where γ_w is the weight matrix with $(\gamma_w)_{st;ij} = w_{st}$, and $\bar{\gamma}_w$ is its vector version.

Taking $\|\cdot\|_{\infty,2}$ of both sides gives, for the elementwise product $A \circ B$ between the same size of matrices A, B,

$$\begin{aligned}&\lambda_n \left\|\bar{\gamma}_w \circ \bar{Z}_{S^c}\right\|_{\infty,2} \\ &= (\left|\left|\left|\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}\right|\right|\right|_{\infty,2} + 1)(\left\|\bar{W}_S\right\|_{\infty,2} + \left\|\bar{R}_S\right\|_{\infty,2}) \\ &\quad + \lambda_n \left\|\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}(\bar{\gamma}_w \circ \bar{Z}_{S^c})\right\|_{\infty,2} \\ &\overset{(a)}{\leq} (\frac{w_{min}}{w_{max}}(1-\alpha) + 1) \cdot 2 \cdot \frac{w_{min}w_{max}\alpha}{4\,(w_{max} + w_{min})}\lambda_n \\ &\quad + w_{max}\left|\left|\left|\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}\right|\right|\right|_{\infty,2}\lambda_n \\ &\leq \lambda_n w_{min}\frac{\alpha}{2} + \lambda_n w_{min}(1-\alpha) \\ &\leq \lambda_n w_{min},\end{aligned}$$

where (a) holds due to the assumptions on $\|W\|_{\infty,2}, \|R\|_{\infty,2}$, and incoherence condition on $\Gamma^*_{S^cS}(\Gamma^*_{SS})^{-1}$.

Based on $\left\|\bar{\gamma}_w \circ \bar{Z}_{S^c}\right\|_{\infty,2} \geq w_{min}\left\|\bar{Z}_{S^c}\right\|_{\infty,2}$ and the inequality above, we have $\left\|\bar{Z}_{S^c}\right\|_{\infty,2} < 1$.

Lemma 5.3 *For the radius* $r := 2\kappa_{\Gamma^*}(\|W\|_{\infty,2} + w_{max}\lambda_n) \leq \min\{1/(3\kappa_{\Sigma^*}d\sqrt{m_{max}}), 1/(3\kappa^3_{\Sigma^*}\kappa_{\Gamma^*}d)\}$*, the error* $\|\Delta\|_{\infty,2}$ *is bounded by* r*.*

Proof. Set the radius $r := 2\kappa_{\Gamma^*}(\|W\|_\infty + \lambda_n w_{max})$ and suppose $r \leq \min\{1/(3\kappa_{\Sigma^*}d\sqrt{m_{max}}), 1/(3\kappa^3_{\Sigma^*}\kappa_{\Gamma^*}d)\}$. From Brouwer's fixed point theorem [19], we can show that it suffices to show $F(\bar{\Delta}_S) \leq r$ for $\|\Delta\|_{\infty,2} \leq r$ where F is the map defined by

$$\begin{aligned}F(\bar{\Delta}_S) = (\Gamma^*_{SS})^{-1}&\mathbf{vec}[(\mathbf{\Theta}^{*-1}\Delta)^2 J\mathbf{\Theta}^{*-1}]_S \\ &- (\Gamma^*_{SS})^{-1}(\bar{W}_S + \lambda_n\bar{\gamma}_S \otimes \bar{Z}_S).\end{aligned}$$

Let T_1 be the first term and T_2 be the second term.

The second term $\|T_2\|_{\infty,2} \leq \kappa_{\Gamma^*}(\|W\|_\infty + \lambda_n w_{max}) = r/2$ by our choice of r at the begining.

The first term

$$\begin{aligned}\|T_1\|_{\infty,2} &\leq \kappa_{\Gamma^*}\left\|\mathbf{vec}[(\mathbf{\Theta}^{*-1}\Delta)^2 J\mathbf{\Theta}^{*-1}]_S\right\|_{\infty,2} \\ &\leq \kappa_{\Gamma^*}\|R(\Delta)\|_{\infty,2} \\ &\overset{(a)}{\leq} \kappa_{\Gamma^*}\frac{3}{2}\kappa^3_{\Sigma^*}\sqrt{m_{max}}d\,\|\Delta\|^2_{\infty,2} \\ &\leq \frac{1}{2}(3\kappa_{\Gamma^*}\kappa^3_{\Sigma^*}d\sqrt{m_{max}} \cdot r)r \\ &\overset{(b)}{\leq} \frac{r}{2},\end{aligned}$$

where (a) holds due to Lemma 4.1 under the assumption $r \leq 1/(3\kappa_{\Sigma^*}d)$, and (b) holds under the assumption on $r \leq \frac{1}{3\kappa^3_{\Sigma^*}\kappa_{\Gamma^*}\sqrt{m_{max}}d}$. Therefore, the error is bounded by $r := 2\kappa_{\Gamma^*}(\|W\|_{\infty,2} + w_{max}\lambda_n) \leq \min\{1/(3\kappa_{\Sigma^*}d), 1/(\kappa^3_{\Sigma^*}\kappa_{\Gamma^*}d)\}$.

Finally, we are ready to present the proof of main theorem.

5.2 Main Proof

Assumptions. We make three assumptions, partly adopted from Loh et al. [15] and Ravikumar et al. [19].

1. The PE-MRF has an underlying graphical structure with singleton separator sets with the maximum degree d.
2. Boundedness condition: The expected value of the sufficient statistic is bounded, and the elementwise absolute value of $\mathbf{Cov}[\boldsymbol{B}(X)]$ is bounded by $\kappa_{\mathbf{Cov}[B]} < \infty$.
3. Incoherence condition: $\left|\left|\left|\Gamma_{S^cS}\,(\Gamma_{S^cS})^{-1}\right|\right|\right|_{\infty,2} \leq \frac{w_{min}}{w_{max}}(1-\alpha)$ for some $\alpha \in (0, 1]$.

Theorem 5.4 *Suppose a PE-MRF* X *satisfies all three assumptions. For a regularization parameter* $\lambda_n > \frac{8(1+w_{min}/w_{max})}{\alpha}\kappa_{\mathbf{Cov}[B]}\sqrt{\frac{\log(m_{max}p)}{n}}$, *let* $\hat{\mathbf{\Theta}}$ *be the unique solution of the group graphical lasso. If the number of samples is given by* $n > c_2(\kappa_{\Sigma^*}, \kappa_{\Gamma^*}, \kappa_{\mathbf{Cov}[B]}, \kappa_{\mathbf{Cov}_{min}}, w_{max}, w_{min}, \alpha)\log(m_{max}p)$, *then the following two statements about* $\hat{\mathbf{\Theta}}$ *hold with probability at least* $1 - e^{-c_1 n}$*:*

1. $\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_{\infty,2} < 2\kappa_{\Sigma^*}\left(\frac{\alpha}{4(1+w_{min}/w_{max})} + w_{max}\right)\lambda_n$ *where* $\mathbf{\Theta}^* = (M_1[\mathbf{E}[B(X)]] + D)^{-1}$.
2. *The recovered edge set* $E(\hat{\mathbf{\Theta}}) = \{(s,t) \mid \left\|\hat{\Theta}_{ij}\right\|_2 > 2\kappa_{\Sigma^*}\left(\frac{\alpha}{4(1+w_{min}/w_{max})} + w_{max}\right)\lambda_n\}$ *becomes the same as the real edge set* $E(\mathbf{\Theta}^{true})$.

Here d is the maximum degree of the graph structure, and $c_2(\kappa_{\Sigma^*}, \kappa_{\Gamma^*}, \kappa_{\mathbf{Cov}[B]}, \kappa_{\mathbf{Cov}_{min}}, w_{max}, w_{min}, \alpha) = 4\kappa_{\Gamma^*}^2 \ (1 + \frac{4(w_{min}+w_{max})}{\alpha})^2 \kappa_{\mathbf{Cov}[B]}^2 \max\{9\kappa_{\Sigma^*}^2 d^2, 9\kappa_{\Sigma^*}^6 \kappa_{\Gamma^*}^2 d^2, 2/\kappa_{\mathbf{Cov}_{min}}\}$.

Proof. Let's set $\delta_n \geq 2m_{max}\sqrt{\kappa_{\mathbf{Cov}[B]}}\sqrt{\frac{\log(m_{max}p)}{n}}$, the event $\mathcal{A} = \{\|W\|_{\infty,2} < \delta_n\}$. Then, by Lemma 4.1, $\mathbf{Pr}\left[\mathcal{A}\right] > 1 - e^{-c_1 n}$.

Now suppose such an event $\mathcal{A}$ occurs. Then, the choice of $\lambda_n = c_3\delta_n$ where $c_3 = \frac{4(w_{max}+w_{min})}{w_{min}w_{max}\alpha}$ satisfies the assumption of Lemma 4.2. Moreover, the error can be expressed with δ_n (or λ_n) as

$$r := 2\kappa_{\Gamma^*}(\|W\|_{\infty,2} + \lambda_n w_{max}) \leq 2\kappa_{\Gamma^*}(1 + c_3 w_{max})\delta_n.$$

We suppose

$$2\kappa_{\Gamma^*}(1 + c_3 w_{max})\delta_n \leq \min\left\{\frac{1}{3\kappa_{\Sigma^*} d}, \frac{1}{3\kappa_{\Sigma^*}^3 \kappa_{\Gamma^*} d}\right\}. \quad (7)$$

This assumption is equivalent to imposing the following restriction on the number of samples n:

$$2\kappa_{\Gamma^*}(1 + c_3 w_{max})\sqrt{\kappa_{\mathbf{Cov}[B]}}\sqrt{\frac{\log(m_{max}p)}{n}} \leq \min\left\{\frac{1}{3\kappa_{\Sigma^*} d}, \frac{1}{3\kappa_{\Sigma^*}^3 \kappa_{\Gamma^*} d}\right\},$$

or

$$n \geq 4\kappa_{\Gamma^*}^2(1 + \frac{4\left(1 + w_{max}/w_{min}\right)}{\alpha})^2 \kappa_{\mathbf{Cov}[B]} \max\left\{9\kappa_{\Sigma^*}^2 d^2, 9\kappa_{\Sigma^*}^6 \kappa_{\Gamma^*}^2 d^2\right\}.$$

Therefore,

$$\begin{aligned} \|R(\Delta)\|_{\infty,2} &\leq \frac{3}{2} d \|\Delta\|_{\infty,2}^2 \kappa_{\Sigma^*}^3 \\ &\overset{(a)}{\leq} \left\{6\kappa_{\Sigma^*}^3 \kappa_{\Gamma^*}^2 (1 + c_3 w_{max})^2 \delta_n\right\} \frac{\lambda_n}{c_3} \\ &\leq \frac{\lambda_n}{c_3}, \end{aligned}$$

where (a) holds due to Lemma 4.1.

Finally, since this choice of $\lambda_n = c_3\delta_n$ satisfies the assumptions on Lemma 4.3 from (7), we conclude that the error $\|\Delta\|_{\infty,2}$ is bounded by $r = 2\kappa_{\Sigma^*}\left(\frac{\alpha}{4(1+w_{min}/w_{max})} + w_{max}\right)\lambda_n$ under the event $\mathcal{A}$. Thus, the statement holds with probability at least $1 - e^{-c_1 n}$.

Now, we move to the second statement. Note that, by the Schur complement, the edge counterpart of $(\mathbf{\Theta}^*)$ is equivalent to the inverse covariance matrix of $b_{node}(X)$, i.e., $\mathbf{\Theta}_{edge}^* = (\mathbf{Cov}[b_{node}(X)])^{-1}$. According to Corollary 2 from Loh et al. [15], which was established only for discrete models but can be extended to PE-MRFs with node-potential $b_{node}(x)$, our generalized inverse covariance matrix $(\mathbf{Cov}[b_{node}(X)])^{-1}$ has the same matrix structure as the real graphical strcture with singleton seperator sets. Therefore, it suffices to estimate $\mathbf{\Theta}^*$ to find the graphical structure of a PE-MRF.

Recall that the first statement demonstrates that the elementwise difference between our estimator $\hat{\mathbf{\Theta}}$ and $\mathbf{\Theta}^*$ is at most r. Therefore, for the minimum value $\kappa_{\mathbf{Cov}_{min}}$ among nonzero absolute elements in $(\mathbf{Cov}[b_{node}(X)])^{-1}$, the graphical structure between $\hat{\mathbf{\Theta}}$ and $\mathbf{\Theta}^*$ matches if $r < \kappa_{\mathbf{Cov}_{min}}/2$. As a result, the recovered edge set $E(\hat{\mathbf{\Theta}}) = \{(s,t) \mid \left\|\hat{\Theta}_{ij}\right\|_2 > 2\kappa_{\Sigma^*}\left(\frac{\alpha}{4(1+w_{min}/w_{max})} + w_{max}\right)\lambda_n\}$ becomes the same as the real edge set $E(\mathbf{\Theta}^{true})$ with probability at least $1 - e^{-c_1 n}$.

References

[1] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[3] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.