Supplementary material for
‘Selection and estimation for mixed graphical models’

BY SHIZHE CHEN, DANIELA WITTEN AND ALI SHOJAIE
Department of Biostatistics, University of Washington, Box 357232, Seattle, Washington 98195, U.S.A.
szchen@uw.edu dwitten@uw.edu ashojaie@uw.edu

S1. TECHNICAL PROOFS OF THE THEORETICAL RESULTS

Proof of Proposition 1

First of all, it is easy to see that if \( \theta_{st} = \theta_{ts} \), then any function \( g \) such that

\[
g(x) \propto \exp \left\{ \sum_{s=1}^{p} f_s(x_s) + \frac{1}{2} \sum_{s=1}^{p} \sum_{t \neq s} \theta_{ts} x_s x_t \right\}
\]

is capable of generating the conditional densities in (3) of the main paper, as long as the function \( g \) is integrable with respect to \( x_s \) for \( s = 1, \ldots, p \). The function \( g \) can be decomposed as

\[
g(x) \propto \exp \left\{ f_s(x_s) + \frac{1}{2} \sum_{t \neq s} (\theta_{ts} + \theta_{st}) x_s x_t \right\} \exp \left\{ \sum_{t \neq s} f_t(x_t) + \frac{1}{2} \sum_{t \neq s, j \neq s, j \neq t} \theta_{tj} x_j x_t \right\},
\]

so the integrability of the conditional density \( p(x_s | x_{-s}) \) guarantees the integrability of \( g \) with respect to \( x_s \). Therefore, the conditional densities of the form (3) in the main paper are compatible if \( \theta_{ts} = \theta_{st} \).

We now prove that any function \( h \) that is capable of generating the conditional density in (3) of the main paper is in the form (S1). The following proof is essentially the same as that in Besag (1974). Suppose \( h \) is a function that is capable of generating the conditional densities. Define

\[
P(x) = \log \{ h(x) / h(0) \},
\]

where 0 can be replaced by any interior point in the sample space.

By definition, \( P(0) = \log \{ h(0) / h(0) \} = 0 \). Therefore, \( P \) can be written in the general form

\[
P(x) = \sum_{s=1}^{p} x_s G_s(x_s) + \sum_{t \neq s} \frac{G_{ts}(x_t, x_s)}{2} x_t x_s + \sum_{t \neq s, t \neq j, j \neq s} \frac{G_{taj}(x_t, x_s, x_j)}{6} x_t x_s x_j + \cdots,
\]

where we express the function \( P \) as the sum of interactions of different orders. Note that the factor of 1/2 is due to \( G_{st}(x_s, x_t) = G_{ts}(x_t, x_s) \); similar factors are present for higher-order interactions. Recalling that we assume \( h \) is capable of generating the conditional density \( p(x_s | x_{-s}) \), from Definition 1 in the main paper we know that

\[
P(x) - P(x^0) = \log \left\{ \frac{h(x) / \int h(x) \, dx}{h(x^0) / \int h(x) \, dx} \right\} = \log \left\{ \frac{p(x_s | x_{-s})}{p(0 | x_{-s})} \right\},
\]
where \( x_0^0 = (x_1, \ldots, x_{s-1}, 0, x_{s+1}, \ldots, x_p)^T \) and \( p(x_s \mid x_{-s}) \) is the conditional density in (3) of the main paper. It follows that

\[
\log \left\{ \frac{p(x_s \mid x_{-s})}{p(0 \mid x_{-s})} \right\} = P(x) - P(x_0^0) = x_s \left( G_s(x_s) + \sum_{t : t \neq s} x_t G_{ts}(x_t, x_s) + \cdots \right). \tag{S2}
\]

Letting \( x_t = 0 \) for \( t \neq s \) in (S2) and using the form of the conditional densities in (3), we have

\[ x_s G_s(x_s) = f_s(x_s) - f_s(0). \tag{S3} \]

Here we set \( f_s(0) = 0 \) since \( f_s(0) \) is a constant. For the second-order interaction \( G_{ts} \), we let \( x_j = 0 \) for \( j \neq t, j \neq s \) in (S2):

\[ x_s G_s(x_s) + x_s x_t G_{ts}(x_t, x_s) = \theta_{st} x_t x_s + f_s(x_s). \]

Similarly, by applying the previous argument to \( P(x) - P(x_t^0) \), we obtain

\[ x_t G_t(x_t) + x_s x_t G_{st}(x_s, x_t) = \theta_{ts} x_t x_s + f_t(x_t). \]

Therefore, if \( \theta_{st} = \theta_{ts} \), then by (S3) we have

\[ G_{st}(x_s, x_t) = G_{ts}(x_t, x_s) = \theta_{st}. \]

It is easy to show that, upon setting \( x_k = 0 \) (where \( k \neq s, k \neq t, k \neq j \)) in (S2), the third-order interactions in \( P(x) \) are zero. Similarly, we can show that fourth- and higher-order interactions are also zero. Hence we arrive at the following formula for \( P \):

\[ P(x) = \sum_{s=1}^p f_s(x_s) + \frac{1}{2} \sum_{s=1}^p \sum_{t \neq s} \theta_{ts} x_s x_t. \]

Furthermore, \( P(x) = \log \{ h(x) \mid h(0) \} \), so the function \( h \) takes the form

\[ h(x) \propto \exp \{ P(x) \} = \exp \left\{ \sum_{s=1}^p f_s(x_s) + \frac{1}{2} \sum_{s=1}^p \sum_{t \neq s} \theta_{ts} x_s x_t \right\}, \]

which is the same as (S1).

**Proof of Lemma 1**

We first prove the claim about compatibility.

It is easy to verify that the conditional densities are integrable given the restrictions with daggers in Table 1 of the main paper. Therefore, these restrictions are sufficient for compatibility.

We now show that the restrictions with daggers in Table 1 are necessary, by investigating each of the distributions given in (4)–(7) of the main paper. Note that we have limited our discussion to the case where all conditional densities are nondegenerate. Recall that we refer to the type of distribution of \( x_s \) given in (4)–(7) of the main paper. Therefore, these restrictions are sufficient for compatibility.

We now show that the restrictions with daggers in Table 1 are necessary, by investigating each of the distributions given in (4)–(7) of the main paper. Note that we have limited our discussion to the case where all conditional densities are nondegenerate. Recall that we refer to the type of distribution of \( x_s \) given in (4)–(7) of the main paper. Therefore, these restrictions are sufficient for compatibility.
Suppose that \( x_s \) is Gaussian, as in (4) of the main paper. Then \( \alpha_{2s} \) has to be negative for the conditional density to be well-defined.

Suppose that \( x_s \) is Bernoulli or Poisson, as in (5) or (6) of the main paper. We can see that there are no restrictions on \( n_s \), and thus no restrictions on \( \theta_{ts} \) or \( \alpha_{1s} \).

Hence, the conditions with daggers in Table 1 are necessary for the conditional densities in (4)–(7) to be compatible.

Next, we show the statement about strong compatibility.

We first prove the necessity of the conditions in Table 1. Recall from Definition 1 that in order for strong compatibility to hold, compatibility must hold, and any function \( g \) that satisfies (8) in the main paper must be integrable. Therefore, we derive the necessary conditions for \( g \) to be integrable.

For Gaussian nodes that are indexed by \( J \), recall that \( \Theta_{JJ} \) is defined as in (9) of the main paper. Then, from properties of the multivariate Gaussian distribution, \( \Theta_{JJ} \) must be negative definite if the joint density exists and is nondegenerate.

Let \( x_1 \) be a Poisson node and \( x_2 \) an exponential node. Consider the ratio

\[
G(x_1, x_2) = \frac{g(x_1, x_2, 0, \ldots, 0)}{g(0, 0, 0, \ldots, 0)} = \exp\{-\log(x_1!) + \alpha_{11}x_1 + \theta_{12}x_1x_2 + \alpha_{12}x_2\},
\]

where \( g \) is the function in (8) of the main paper. It is not hard to see that integrability of \( G(x_1, x_2) \) is a necessary condition for integrability of the joint density. Summing over \( x_1 \) yields

\[
\sum_{i=0}^{\infty} G(i, x_2) = \exp\{\alpha_{12}x_2 + \exp(\alpha_{11} + \theta_{12}x_2)\}.
\]

Therefore, if \( \sum_{i=0}^{\infty} G(i, x_2) \) is integrable with respect to the exponential node \( x_2 \), it must be the case that \( \theta_{12} = \theta_{21} \leq 0 \). Following a similar argument, the edge potential \( \theta_{12} = \theta_{21} \) has to be nonpositive when \( x_2 \) is Poisson and zero when \( x_2 \) is Gaussian.

A similar argument to the one just described can be applied to the exponential nodes. Such an argument reveals that conditions on the edge potentials of the exponential nodes that are necessary for \( g \) to be a density are those stated in Table 1 of the main paper.

For Bernoulli nodes, no restrictions on the edge potentials are necessary in order for \( g \) to be a density.

Therefore, the conditions listed in Table 1 are necessary for the conditional densities in (4)–(7) to be strongly compatible.

Finally, we show that the conditions listed in Table 1 are sufficient for the conditional densities to be strongly compatible. We can restrict the discussion by conditioning on the Bernoulli nodes, since integrating over Bernoulli variables yields a mixture of finite components. Table 1 guarantees that the Gaussian nodes are isolated from the Poisson and exponential nodes, as the corresponding edge potentials are zero. From Table 1, the distribution of Gaussian nodes is integrable, as \( \Theta_{JJ} \) in (9) is negative definite. We now consider the Poisson and exponential nodes. For these,

\[
\exp\left\{ \sum_{s=1}^{p} f_s(x_s) + \frac{1}{2} \sum_{s=1}^{p} \sum_{t \neq s} \theta_{st}x_sx_t \right\} \leq \exp\left\{ \sum_{s=1}^{p} f_s(x_s) \right\}
\]

since \( \theta_{st}x_sx_t \leq 0 \). So the joint density is dominated by the density of a model with no interactions, which is integrable since \( \alpha_{1t} \) for an exponential node \( x_t \) is nonpositive; this follows
from the fact that $0 \leq \sum_{s \in I} |\theta_{st}| < -\alpha_{1i}$, as stated in Table 1. Therefore, the conditions listed in Table 1 are also sufficient for the conditional densities in (4)–(7) to be strongly compatible.

Proof of Theorem 1

Our proof is similar to that of Theorem 1 in Yang et al. (2012), and is based on the primal-dual witness method (Wainwright, 2009). The primal-dual witness method studies the properties of $\ell_1$-penalized estimators by investigating the subgradient condition of an oracle estimator. We assume that readers are familiar with the primal-dual witness method; for reference, see Ravikumar et al. (2011) and Yang et al. (2012). Without loss of generality, we assume $s = p$ to avoid cumbersome notation. For other values of $s$, a similar proof holds with more complicated notation. Below we write $\Theta_p$ as $\theta$, $\eta_p$ as $\eta$, and $\ell_p$ as $\ell$ for simplicity. We also denote the neighbours of $x_p$, $N(x_p)$, by simply $N$.

The subgradient condition for (10) in the main paper with respect to $(\theta^T, \alpha_{1p})^T$ is
\[
-\nabla \ell(\theta, \alpha_{1p}; X) + \lambda_n Z = 0, \quad Z_t = \text{sgn}(\theta_t) \text{ for } t < p, \quad Z_p = 0,
\] (S4)
where
\[
\text{sgn}(x) = \begin{cases} x/|x|, & x \neq 0, \\ \gamma \in [-1, 1], & x = 0. \end{cases}
\]

We construct the oracle estimator $(\hat{\theta}_N^T, \hat{\alpha}_N, \hat{\alpha}_{1p})^T$ as follows: first, let $\hat{\theta}_N = 0$ where $\Delta$ indicates the set of non-neighbours; second, obtain $\hat{\theta}_N$ and $\hat{\alpha}_{1p}$ by solving (10) in the main paper with the additional restriction that $\hat{\theta}_N = 0$; third, set $\hat{Z}_t = \text{sgn}(\hat{\theta}_t)$ for $t \in N$ and $\hat{Z}_p = 0$; last, estimate $\hat{Z}_\Delta$ from (S4) by plugging in $\hat{\theta}, \hat{\alpha}_{1p}$ and $\hat{Z}_\Delta$. To complete the proof, we verify that $(\hat{\theta}_N^T, \hat{\alpha}_N, \hat{\alpha}_{1p})^T$ and $\hat{Z} = (\hat{Z}_N^T, \hat{Z}_\Delta^T, 0)^T$ form a primal-dual pair of (10) which recovers the true neighbourhood exactly.

Applying the mean value theorem to each element of $\nabla \ell(\hat{\theta}, \hat{\alpha}_{1p}; X)$ in the subgradient condition (S4) gives
\[
Q^*(\frac{\hat{\theta} - \theta^*}{\hat{\alpha}_{1p} - \alpha_{1p}^*}) = -\lambda_n \hat{Z} + W^n + R^n,
\] (S5)
where $W^n = \nabla \ell(\theta^*, \alpha_{1p}^*; X)$ is the sample score function evaluated at the true parameter $(\theta^*^T, \alpha_{1p}^*)^T$. Recall that $Q^* = -\nabla^2 \ell(\theta^*, \alpha_{1p}^*; X)$ is the negative Hessian of $\ell(\theta, \alpha_{1p}; X)$ with respect to $(\theta^T, \alpha_{1p})^T$, evaluated at the true values of the parameters. In (S5), $R^n$ is the residual term from the mean value theorem, whose $k$th term is
\[
R_k^n = [\nabla^2 \ell(\hat{\theta}^k, \hat{\alpha}_{1p}^k; X) - \nabla^2 \ell(\theta^*, \alpha_{1p}^*; X)]_k^T \left(\frac{\hat{\theta} - \theta^*}{\hat{\alpha}_{1p} - \alpha_{1p}^*}\right),
\] (S6)
where $\hat{\theta}^k$ denotes an intermediate point between $\theta^*$ and $\hat{\theta}$, $\hat{\alpha}_{1p}^k$ denotes an intermediate point between $\alpha_{1p}^*$ and $\hat{\alpha}_{1p}$, and $[\cdot]_k^T$ denotes the $k$th row of a matrix.

By construction, $\hat{\theta}_N = 0$. Thus, (S5) can be rearranged as
\[
\lambda_n \hat{Z}_\Delta = (W^n + R^n_{\Delta}) - Q^*_{\Delta^c}(Q^*_{\Delta^c})^{-1}(W^n_{\Delta^c} + R^n_{\Delta^c} - \lambda_n \hat{Z}_{\Delta^c}).
\] (S7)
We obtain an estimator $\hat{Z}_\Delta$ by plugging $\hat{\theta}, \hat{\alpha}_{1p}$ and $\hat{Z}_{\Delta^c}$ into (S7). To complete the proof, we need to verify strict dual feasibility,
\[
\|\hat{Z}_\Delta\|_\infty < 1,
\]
and sign consistency,
\[ \text{sgn}(\hat{\theta}_t) = \text{sgn}(\theta^*_t) \quad \text{for any } t \in N. \]

In (S7), \( \max_{t \in \Delta} \|Q^*_t (Q^*_\Delta e)^{-1}\|_1 \leq 1 - a \) by Assumption 1 in the main paper. The following lemmas characterize useful concentration inequalities regarding \( W^m, R^m \) and \( \hat{\theta}_N - \theta^*_N \). Proofs of Lemmas S1 and S2 are given in the next two subsections.

**Lemma S1.** Suppose that
\[ \frac{8(2 - a)}{a} \{ \delta_2 \kappa_2 \log(2p) / n \}^{1/2} \leq \lambda_n \leq \frac{2(2 - a)}{a} \delta_2 \kappa_2 M, \]
where \( \delta_2 \) is defined in Proposition 3 and \( a \) and \( \kappa_2 \) are defined, respectively, in Assumptions 1 and 3 of the main paper. Then
\[ \text{pr} \left( \|W^n\|_\infty > \frac{a \lambda_n}{8 - 4a} \right) \leq \exp(-c_3 \delta_3 n), \]
where \( \delta_3 = 1/(\kappa_2 \delta_2) \) and \( c_3 \) is some positive constant.

**Lemma S2.** Suppose that \( \xi_1 \) and \( \|W^n\|_\infty \leq a \lambda_n / (8 - 4a) \) hold and that
\[ \lambda_n \leq \min \left\{ \frac{a \Lambda_1^* (d + 1)^{-1}}{288(2 - a) \kappa_2 \Lambda_2}, \frac{a \Lambda_2^* (d + 1)^{-1}}{12 \Lambda_2 \kappa_3 \delta_1 \log p} \right\}, \]
where \( \delta_1 \) is defined in Proposition 2 and \( a \) and \( \kappa_3 \) are defined, respectively, in Assumptions 1 and 3 of the main paper. Then, with probability 1,
\[ \|\hat{\theta}_N - \theta^*_N\|_2 < \frac{10}{\Lambda_1} (d + 1)^{1/2} \lambda_n, \quad \|R^n\|_\infty \leq \frac{a \lambda_n}{8 - 4a}. \]

We now continue with the proof of Theorem 1. Given Assumption 6, the conditions regarding \( \lambda_n \) are met for Lemmas S1 and S2.

We now assume that \( \xi_1, \xi_2 \) and the event \( \|W^n\|_\infty \leq a \lambda_n / (8 - 4a) \) are true, so that the conditions for the two lemmas are satisfied. We derive the lower bound for the probability of these events at the end of the proof.

First, applying Lemma S2 and Assumption 1 to (S7) yields
\[ \|Z_\Delta\|_\infty \leq \max_{t \in \Delta} \|Q^*_t (Q^*_\Delta e)^{-1}\|_1 \left( \|W^n\|_\infty + \|R^n\|_\infty + \lambda_n \|Z_\Delta\|_\infty \right) / \lambda_n \]
\[ + \left( \|W^n\|_\infty + \|R^n\|_\infty \right) / \lambda_n \]
\[ \leq (1 - a) + (2 - a) \left\{ \frac{a}{4(2 - a)} + \frac{a}{4(2 - a)} \right\} < 1. \]  \( \text{(S8)} \)

Next, applying Lemma S2 and a norm inequality to \( \|\hat{\theta}_N - \theta^*_N\|_\infty \) gives
\[ \|\hat{\theta}_N - \theta^*_N\|_\infty \leq \|\hat{\theta}_N - \theta^*_N\|_2 < \frac{10}{\Lambda_1} (d + 1)^{1/2} \lambda_n \leq \min_t |\theta_t|, \]  \( \text{(S9)} \)
since \( \min_t |\theta_t| \geq 10(d + 1)^{1/2} \lambda_n / \Lambda_1 \) by Assumption 5. The strict inequality in (S9) ensures that the sign of the estimator is consistent with the sign of the true value for all edges.

Inequalities (S8) and (S9) are sufficient to establish the result, i.e., \( \hat{N} = N \). Let \( A \) be the event \( \|W^n\|_\infty \leq a \lambda_n / (8 - 4a) \). Recall that we have assumed events \( A, \xi_1 \) and \( \xi_2 \) to be true in order to prove (S8) and (S9). We now derive the lower bound for the probability of \( A \cap \xi_1 \cap \xi_2 \).
Using the fact that
\[ \Pr\{A \cap \xi_1 \cap \xi_2 \cap \xi_3 \} \leq \Pr(A^c \mid \xi_1 \cap \xi_2) + \Pr(\xi_1 \cap \xi_2) \leq \Pr(A^c \mid \xi_1, \xi_2) + \Pr(\xi_1) + \Pr(\xi_2), \]
we know that the probability of \( A \cap \xi_1 \cap \xi_2 \) satisfies
\[ \Pr\left( \|W^n\|_\infty \leq \frac{a}{2} - \frac{\lambda_n}{4} \right) \cap \xi_2 \cap \xi_3 \geq 1 - c_1 p^{-\delta_1 + 2} - \exp(-c_2 \delta_2 n) - \exp(-c_3 \delta_3 n), \]
where \( c_1, c_2 \) and \( c_3 \) are the constants from Proposition 2, Proposition 3 and Lemma S1. Thus, the event \( A \cap \xi_1 \cap \xi_2 \) occurs with high probability when the sample size \( n \) is large. This completes the proof.

**Proof of Lemma S1**

Recall that \( \eta^{(i)} = \alpha_{1p} + \sum_{t < p} \theta_t x_t^{(i)} \) and that we have assumed \( \alpha_{kp} \) is known for \( k \geq 2 \). We can rewrite the conditional density in (3) of the main paper as
\[ p(x_p \mid x_{-p}) \propto \exp\{\eta x_p - D(\eta)\}. \]
For any \( t < p \),
\[ W_t^n = \frac{\partial \ell}{\partial \theta_t} = n \frac{\partial \ell}{\partial \eta^{(i)}} \frac{\partial \eta^{(i)}}{\partial \theta_t} = \frac{n}{n} \sum_{i=1}^n \{x_p^{(i)} - D'(\eta^{(i)})\} x_t^{(i)}. \quad (S10) \]

Recall that \( M \) is a large constant introduced in Assumption 3. Suppose that \( M \) is large enough that \( |\alpha_{1p}^* + \sum_{k < p} |\theta_k^*| < M/2 \). For every \( v \) such that \( 0 < v < M/2 \),
\[
E \left( \exp \left[ vx_t^{(i)} \left\{ x_p^{(i)} - D'(\eta^{(i)}) \right\} \right] \bigg| X_{-p} \right) \\
= E \left\{ \exp \left[ vx_t^{(i)} x_p^{(i)} \bigg| X_{-p} \right] \exp \left\{ -vx_t^{(i)} D'(\eta^{(i)}) \right\} \right\} \\
= \exp \left\{ D(\eta^{(i)} + vx_t^{(i)}) - D(\eta^{(i)}) \right\} \exp \left\{ -vx_t^{(i)} D'(\eta^{(i)}) \right\} \\
= \exp \left\{ vx_t^{(i)} D'(\eta^{(i)}) + (vx_t^{(i)})^2 \frac{D''(\eta)}{2} \right\} \exp \left\{ -vx_t^{(i)} D'(\eta^{(i)}) \right\} \\
= \exp \left\{ \left(vx_t^{(i)}\right)^2 \frac{D''(\tilde{\eta})}{2} \right\}, \quad \tilde{\eta} \in [\eta^{(i)}, \eta^{(i)} + vx_t^{(i)}], \quad (S11) \]
where the second equality was derived using properties of the moment generating function of the exponential family, and the third equality follows from a second-order Taylor expansion. Since \( \tilde{\eta} \in [\eta^{(i)}, \eta^{(i)} + vx_t^{(i)}] \), the event \( \xi_1 \) implies that
\[ |\tilde{\eta}| \leq |\alpha_{1p}^*| + \sum_{k < p} |x_k^{(i)} \theta_k^*| + |vx_t^{(i)}| \leq |\alpha_{1p}^*| + \left( \sum_{k < p} |\theta_k^*| + |v| \right) \max_{i,j} |x_t^{(i)}| \leq M \delta_1 \log p. \quad (S12) \]
Therefore, the condition of Assumption 3 is satisfied, and thus $|D^\prime(\tilde{\eta})| \leq \kappa_2$. Recalling that $(x_i^{(i)})_{i=1}^n$ are independent samples, it follows from (S10) and (S11) that

$$
E\{\exp(vnW_t^n) \mid \xi_2, \xi_1\} = E\left[E\{\exp(vnW_t^n) \mid X_p, \xi_2, \xi_1\} \mid \xi_2, \xi_1\right] \\
\leq E\left[E\left\{\exp\left\{v^2\kappa_2 \sum_{i=1}^n (x_i^{(i)})^2\right\} \mid \xi_2, \xi_1\right\}\right] \\
\leq \exp(nv^2\kappa_2^2/2), \quad (S13)
$$

where we have used the event $\xi_2$ in the last inequality. Similarly,

$$
E\{\exp(-vnW_t^n) \mid \xi_2, \xi_1\} \leq \exp(nv^2\kappa_2^2/2). \quad (S14)
$$

Furthermore, one can see from an argument similar to the one for (S10) and (S11) that

$$
E\{\exp(nvW_t^n) \mid \xi_1\} = E\left\{\exp\left(vn\frac{\partial \ell}{\partial \alpha_{1p}}\right) \right\} \\
= \prod_{i=1}^n E\left\{\exp[v\{x_p^{(i)} - D'(\eta^{(i)})\}] \mid \xi_1\right\} \leq \exp(n\kappa_2v^2/2).
$$

We focus on discussion of (S13) and (S14) since $\delta_2 \geq 1$. For some $\delta$ to be specified, we let $v = \delta/(\kappa_2 \delta_2)$ and apply the Chernoff bound (Chernoff, 1952; Ravikumar & Lafferty, 2004) with (S13) and (S14) to get

$$
\Pr(|W_t^n| > \delta \mid \xi_2, \xi_1) \leq \frac{E\{\exp(vnW_t^n) \mid \xi_2, \xi_1\}}{E\{\exp(vn\delta)\}} + \frac{E\{\exp(-vnW_t^n) \mid \xi_2, \xi_1\}}{E\{\exp(vn\delta)\}} \\
\leq 2 \exp\left(-\frac{n \delta^2}{2\kappa_2 \delta_2}\right). \quad (S15)
$$

Upon letting $\delta = a\lambda_n/(8 - 4a)$ and using the Bonferroni inequality, we get

$$
\Pr\left(\|W^n\|_\infty > \frac{a}{2 - a} \frac{\lambda_n}{4} \mid \xi_2, \xi_1\right) \leq 2 \exp\left(-\frac{n a^2 \lambda_n^2}{32(2 - a)^2 \kappa_2 \delta_2} + \log(p)\right) \\
\leq \exp\left(-\frac{a^2 \lambda_n^2}{64(2 - a)^2 \kappa_2 \delta_2} n\right) = \exp(-c_3 \delta_3 n), \quad (S15)
$$

where $\delta_3 = 1/(\kappa_2 \delta_2)$ and $c_3 = a^2 \lambda_n^2/(64(2 - a)^2)$. In (S15), we have made use of the assumption that $\lambda_n \geq 8(2 - a)\kappa_2 \delta_2 \log(2p)/n^{1/2}/a$, and we also require that $\lambda_n \leq 2(2 - a)\kappa_2 \delta_2 M/a$ since $v = a\lambda_n/\{(8 - 4a)\kappa_2 \delta_2\} \leq M/2$.

**Proof of Lemma S2**

We first show that $\|\hat{\theta}_N - \theta_N\|_2 < 10(d + 1)^{1/2}\lambda_n/\Lambda_1$.

Following the method used in Fan & Li (2004) and Ravikumar et al. (2010), we construct a function $F(u)$ as

$$
F(u) = -\ell(\theta^* + u_{-p}; \alpha_{1p}^* + u_p; X) + \ell(\theta^*; \alpha_{1p}^*; X) + \lambda_n\|\theta^* + u_{-p}\|_1 - \lambda_n\|\theta^*\|_1, \quad (S16)
$$

where $u$ is a $p$-dimensional vector and $u_\Delta = 0$. This function has some nice properties: (i) $F(0) = 0$ by definition; (ii) $F(u)$ is convex in $u$ given the form of (3) in the main paper; and (iii) by the construction of the oracle estimator $\hat{\theta}$, $F(u)$ is minimized by $\hat{u}$ with $\hat{u}_{-p} = \hat{\theta} - \theta^*$ and $\hat{u}_p = \hat{\alpha}_{1p} - \alpha_{1p}^*$. 
We claim that if there exists a constant $B$ such that $F(u) > 0$ for any $u$ for which $\|u\|_2 = B$ and $u_\Delta = 0$, then $\|\hat{u}\|_2 \leq B$. To show this, suppose that $\|\hat{u}\|_2 > B$ for such a constant. Let $t = B/\|\hat{u}\|_2$. Then $t < 1$, and the convexity of $F(u)$ gives

$$F(t\hat{u}) \leq (1-t)F(0) + tF(\hat{u}) \leq 0.$$  

Thus $\|t\hat{u}\|_2 = B$ and $(t\hat{u})_\Delta = t\hat{u}_\Delta = 0$, but $F(t\hat{u}) \leq 0$, which is a contradiction.

Applying a Taylor expansion to the first term of $F(u)$ gives

$$F(u) = -\nabla \ell(\theta^*, \alpha_{1p}; X)\, u - u^T \nabla^2 \ell(\theta^* + vu_{-p}, \alpha_{1p} + vu_p; X)u/2 + \lambda_n(\|\theta^* + u_{-p}\|_1 - \|\theta^*\|_1) = I + II/2 + III$$

for some $v \in [0, 1]$. Recall that $u_\Delta = 0$ is defined in (S16). The gradient and Hessian are with respect to the vector $(\theta^T, \alpha_{1p})^T$.

We now proceed to find a $B$ such that for $\|u\|_2 = B$ and $u_\Delta = 0$, the function $F(u)$ is always greater than 0. First, given $\|W^n\|_\infty \leq a\lambda_n/(8 - 4a)$ and $a < 1$ assumed in Assumption 1,

$$|I| = |(W^n)^T u| \leq \|W^n\|_\infty \|u\|_1 \leq \frac{a}{2 - a}\frac{\lambda_n}{4}(d + 1)^{1/2}B \leq \frac{\lambda_n}{4}(d + 1)^{1/2}B.$$  

Next, by the triangle inequality and the Cauchy–Schwarz inequality,

$$III \geq -\lambda_n\|u_{-p}\|_1 \geq -\lambda_n d^{1/2}\|u_{-p}\|_2 \geq -\lambda_n(d + 1)^{1/2}B.$$  

To bound $II$, we observe that

$$-\nabla^2 \ell(\theta^* + vu_{-p}, \alpha_{1p}^* + vu_p; X) = \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^T D''(\eta_r^{(i)}),$$

where $x_0 = (x_{-p}^T, 1)^T$ is as in Assumption 2 and $\eta_r^{(i)} = \alpha_{1p}^* + vu_p + \sum_{t < p}(\theta_t^* + vu_t)x_t^{(i)}$. Applying a Taylor expansion to each $D''(\eta_r^{(i)})$ at $\eta^{(i)} = \alpha_{1p}^* + \sum_{t < p}\theta_t^*x_t^{(i)}$, we get

$$-\nabla^2 \ell(\theta^* + vu_{-p}, \alpha_{1p}^* + vu_p; X) = \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^T D''(\eta^{(i)}) + \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^T D''(\tilde{\eta}^{(i)})(vu^T x_0^{(i)})$$

$$= Q^* + \frac{1}{n} \sum_{i=1}^n x_0^{(i)} (x_0^{(i)})^T D''(\tilde{\eta}^{(i)})(vu^T x_0^{(i)}),$$

where $\tilde{\eta}^{(i)} \in [\eta^{(i)}, \eta_r^{(i)}]$. Using the argument for $\tilde{\eta}$ in (S12) and the facts that $v \leq 1$ and $\|u\|_2 = B$, we can see that $\tilde{\eta}^{(i)}$ is in the range required for Assumption 3 to hold given $\xi_1$. Therefore,
Thus, if our choice of $B$ satisfies
\[ \Lambda_1 - \delta_1 \log(p)B < \frac{\Lambda_1}{2}, \tag{S17} \]
then a lower bound on $F(u)$ is
\[ F(u) \geq -\frac{\lambda_n}{4}(d+1)^{1/2}B + \frac{\Lambda_1}{4}B^2 - \lambda_n(d+1)^{1/2}B. \]
So $F(u) > 0$ for any $B > 5(d+1)^{1/2}\lambda_n/\Lambda_1$. We can therefore take
\[ B = 6(d+1)^{1/2}\lambda_n/\Lambda_1 \tag{S18} \]
to get
\[ \|\hat{\theta}_N - \theta_N^*\|_2 \leq \|\hat{u}\|_2 \leq B = \frac{6}{\Lambda_1}(d+1)^{1/2}\lambda_n. \tag{S19} \]
Thus, $\|\hat{\theta}_N - \theta_N^*\|_2 < 10\lambda_n(d+1)^{1/2}/\Lambda_1$. It is easy to show that (S18) satisfies (S17) provided that
\[ \lambda_n \leq \frac{\Lambda^2_2(d+1)^{-1}}{12\Lambda_2\kappa_3\delta_1 \log p}. \]
To find the bound for $R^m$ defined in (S6), we first recall that $(\bar{\theta}_N, \bar{\alpha})^T$ is an intermediate point between $(\theta^T, \alpha^T_1)$ and $(\bar{\theta}_{N-1}, \bar{\alpha}_{N-1})^T$. We write $\bar{\eta}^i = \bar{\alpha}_p + \sum_{t<p} \bar{\theta}_t x_t^{(i)}$ and observe that $|\bar{\eta}^{(i)}| \leq M\delta_1 \log p$ for $i = 1, \ldots, n$ using the argument of (S12), which implies that Assump-
tion 3 is applicable. Thus,

\[
\Lambda_{\text{max}} \left\{ \nabla^2 \ell(\tilde{\theta}, \tilde{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha^*_{1p}; X) \right\} \\
= \max_{\|u\|_2 = 1} u^T \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( D''(\eta^{(i)}) - D''(\eta^*) \right) x_0^{(i)} (x_0^{(i)})^T \right\} u.
\]

By Assumption 3, \(|D''(\eta^{(i)}) - D''(\eta^*)| \leq 2\kappa_2\), and so

\[
\Lambda_{\text{max}} \left\{ \nabla^2 \ell(\tilde{\theta}, \tilde{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha^*_{1p}; X) \right\} \\
= \max_{\|u\|_2 = 1} u^T \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( D''(\eta^{(i)}) - D''(\eta^*) \right) x_0^{(i)} (x_0^{(i)})^T \right\} u \\
\leq 2\kappa_2 \max_{\|u\|_2 = 1} u^T \left\{ \frac{1}{n} \sum_{i=1}^{n} x_0^{(i)} (x_0^{(i)})^T \right\} u \leq 2\kappa_2 \Lambda_2,
\]

where we have used Assumption 2 for the last inequality. Hence we arrive at

\[
\| R^n \|_\infty \leq \| R^n \|_2^2 = \left\| \left( \nabla^2 \ell(\hat{\theta}, \hat{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha^*_{1p}; X) \right) \right\|^2_2 (\hat{\theta} - \theta^*)^T (\hat{\alpha}_{1p} - \alpha^*_{1p}) \leq \Lambda_{\text{max}} \left\{ \nabla^2 \ell(\tilde{\theta}, \tilde{\alpha}_{1p}; X) - \nabla^2 \ell(\theta^*, \alpha^*_{1p}; X) \right\} \left\| \hat{\theta} - \theta^* \right\|_2^2 \leq 72\kappa_2 \Lambda_2 (d + 1) \lambda_n^2,
\]

where the last inequality follows from (S19). So \( \| R^n \|_\infty \leq a\lambda_n / (8 - 4\alpha) \) if

\[
\lambda_n \leq \frac{a}{2 - a} \frac{\Lambda_1^2}{288(d + 1)\kappa_2 \Lambda_2},
\]

which holds by assumption.

**Proof of Corollary 1**

The proof is essentially the same as the proof of Theorem 1. We first show that a modified version of Lemma S2 holds with fewer conditions.

**Lemma S3.** Suppose that \( p(x_p | x_{-p}) \) follows a Gaussian distribution as in (4) of the main paper and that \( \| W^n \|_\infty \leq a\lambda_n / (8 - 4\alpha) \). Then

\[
\| \hat{\theta}_N - \theta^*_N \|_2 < \frac{10}{\Lambda_1} (d + 1)^{1/2} \lambda_n, \quad \| R^n \|_\infty = 0.
\]
Proof. To prove this lemma, we go through essentially the same argument as in the proof of Lemma S2, but for II we note that

\[ II \geq \min_{u: \|u\|_2 = B, u_\Delta = 0} \left\{ -u^T \nabla^2 \ell(\theta^* + vu_{-p}, \alpha^*_p + vu_p; X)u \right\} \]

\[ \geq B^2 \Lambda_{\min}(-Q_{\Delta e \Delta e}) - \max_{v \in [0,1]} \max_{u: \|u\|_2 = B, u_\Delta = 0} u^T \left\{ \frac{1}{n} \sum_{i=1}^n D''(\tilde{\eta}^{(i)})(vu_{-p}^{(i)}x_0^{(i)}x_0^{(i)T})u \right\} \]

\[ \geq \Lambda_1 B^2 - 0, \]

since \( D''(\tilde{\eta}^{(i)}) = 0 \) for a Gaussian distribution. Hence

\[ F(u) \geq -\frac{\lambda_n}{4} (d + 1)^{1/2} B + \frac{1}{2} \Lambda_1 B^2 - \frac{\lambda_n}{4} (d + 1)^{1/2} B. \]

So \( F(u) > 0 \) for \( B > 5\lambda_n(d + 1)^{1/2}/(2\Lambda_1) \). We can therefore take \( B = 5(d + 1)^{1/2}\lambda_n/\Lambda_1 \) to get

\[ \|\tilde{\theta}_N - \theta^*_N\|_2 \leq \|\tilde{u}\|_2 \leq B = \frac{5}{\Lambda_1} (d + 1)^{1/2}\lambda_n. \]

Thus, \( \|\tilde{\theta}_N - \theta^*_N\|_2 < 10\lambda_n(d + 1)^{1/2}/\Lambda_1; \) also, \( \|R^n\|_{\infty} = 0 \) trivially, as \( D''(\tilde{\eta}^{(i)}) - D''(\eta^*) = 0 \) for a Gaussian distribution.

With Lemma S3, we can then verify (S8) and (S9) as in the proof of Theorem 1. Finally, we drop the requirement of \( \xi_1 \) in the condition of Lemma S3, so the probability of \( N = N \) is

\[ \Pr \left\{ \left( \|W^n\|_{\infty} \leq \frac{a}{2 - a} \frac{\lambda_n}{4} \right) \cap \xi_2 \right\} \geq 1 - \exp(-c_2\delta^2 n) - \exp(-c_3\delta_3n), \]

where \( c_2 \) and \( c_3 \) are constants from Proposition 3 in the main paper and Lemma S1.

S2. Additional Details of the Data-Generation Procedure

Here we provide additional details of the data-generation procedure described in § 6.1 of the main paper. In particular, we describe the approach used to guarantee that the conditions listed in Table 1 for strong compatibility of the conditional distributions are satisfied.

Recall from Table 1 that in order for strong compatibility to hold, the matrix \( \Theta_{JJ} \) in (9) of the main paper that contains the edge potentials between the Gaussian nodes must be negative definite. If \( \Theta_{JJ} \) generated as described in § 6.1 is not negative definite, then we define a matrix \( T_{JJ} \) by

\[ T_{JJ} = -\Theta_{JJ} + \{\Lambda_{\min}(\Theta_{JJ}) - 0.1\} I, \]

where \( \Lambda_{\min}(\Theta_{JJ}) \) denotes the minimum eigenvalue of \( \Theta_{JJ} \). Thus, \( T_{JJ} \) is guaranteed to be negative definite, as all its eigenvalues are no larger than \(-0.1\). We then standardize \( T_{JJ} \) so that its diagonal elements equal \(-1:\)

\[ \tilde{T}_{JJ} = \text{diag}(|T_{11}|^{-1/2}, \ldots, |T_{mm}|^{-1/2}) T_{JJ} \text{diag}(|T_{11}|^{-1/2}, \ldots, |T_{mm}|^{-1/2}). \]

Finally, we replace \( \Theta_{JJ} \) with \( \tilde{T}_{JJ} \).

Table 1 also indicates that for strong compatibility to hold, the edge potential between two Poisson nodes must be negative. Therefore, after generating edge potentials as described in § 6.1, we replace \( \theta_{st} \) with \(-|\theta_{st}| \) where \( x_s \) and \( x_t \) are Poisson nodes.
REFERENCES


