Active Learning Algorithms for Graphical Model Selection: Supplementary Material

Gautam Dasarathy[†], **Aarti Singh**[†], **Maria F. Balcan**[†,*], **and Jong H. Park**[*]

[†]Machine Learning Department
[*]Computer Science Department
Carnegie Mellon University

A Proof of Theorem 1

We will restate the theorem here for convenience.

Theorem 1. *Fix $\delta \in (0,1)$. For each $\ell \leq d_{\max}$, assume that the subroutines* **nbdSelect** *and* **nbdVerify** *satisfy:*

(C1) For any vertex $i \in [p]$ and subset $F \subseteq [p]$ that are such that $|N(i)| = d_i \leq \ell$ and $\overline{N}(i) \subseteq F$, the following holds. Given $g(\ell)$ samples from X_F, **nbdSelect**$(i, \ell, \left\{X_F^{(j)}\right\}_{j \in S_1})$ *returns the true neighborhood of i with probability greater than $1 - \delta/2pd_{\max}$.*

(C2) For any vertex $i \in [p]$ and subsets $F, H \subseteq [p]$ that are such that $|N(i)| = d_i \leq \ell$, $\overline{N}(i) \subseteq F$, and $H \subseteq F$, the following holds. Given $h(|H|)$ samples from X_F, **nbdVerify** $(i, H, \left\{X_F^{(j)}\right\})$ *returns* **true** *if and only if $N(i) \subseteq H$ with probability greater than $1 - \delta/2pd_{\max}$.*

Then, with probability no less than $1 - \delta$, Algorithm 1 returns the correct graph. Furthermore, it suffices if $B \geq \sum_{i \in [p]} \sum_{0 \leq k \leq \lceil \log_2 d^i_{\max} \rceil} g(2^k) + h(2^k))$. That is, Algorithm 1 has a total sample complexity of $\sum_{i \in [p]} \sum_{0 \leq k \leq \lceil \log_2 d^i_{\max} \rceil} g(2^k) + h(2^k)$ at confidence level $1 - \delta$.

Proof. To prove this theorem, we will use a simple argument that can be thought of as a proof by probabilistic induction. Towards this end, we will let $\mathcal{E}_k$ be the event that Algorithm 1 succeeds at iteration number k. Notice that k takes values in the set $\{1, 2, \ldots, \lfloor \log_2(2p) \rfloor\}$ since the algorithm terminates when the (doubling) counter satisfies $\ell = 2^{k-1} \geq 2p$. We can characterize the event $\mathcal{E}_k$ as follows:

- For each $i \in [p] \setminus$ NBDFOUND, if $d_i = |N(i)| \leq 2^{k-1}$, then $\widehat{N}(i)$, the output of `nbdSelect`$(i, \ell, \{X^{(j)}_{[p]\setminus\mathrm{SETTLED}}\}_{j \in S_1})$ is exactly $N(i)$, and `nbdVerify`$(i, \widehat{N}(i), \left\{X_{[p]\setminus\mathrm{SETTLED}}\right\}_{j \in S_2})$ outputs **true**.
- For each $i \in [p] \setminus$ NBDFOUND, if $d_i > 2^{k-1}$, then `nbdVerify`$(i, \widehat{N}(i), \left\{X_{[p]\setminus\mathrm{SETTLED}}\right\}_{j \in S_2})$ outputs **false**.

We begin our proof by bounding the probability of error from above as follows.

$$\begin{aligned} \mathbb{P}\left[\text{error}\right] &= \mathbb{P}\left[\bigcup_{k=1}^{\lfloor \log_2(2p) \rfloor} \mathcal{E}_k^c\right] \\ &\leq \sum_{k=1}^{\lfloor \log_2(2p) \rfloor} \mathbb{P}\left[\mathcal{E}_k^c | \mathcal{E}_1, \ldots, \mathcal{E}_{k-1}\right], \end{aligned} \tag{1}$$

where we have used the convention that $\mathcal{E}_1, \ldots, \mathcal{E}_{k-1} = \emptyset$ for $k = 1$. In what follows, fix an arbitrary $k \in [\lfloor \log_2(2p) \rfloor]$ and set $\ell = 2^{k-1}$; we will bound the probability $\mathbb{P}\left[\mathcal{E}_k^c | \mathcal{E}_1, \ldots, \mathcal{E}_{k-1}\right]$.

First, observe that conditioned on $\mathcal{E}_1, \ldots, \mathcal{E}_{k-1}$, the following is true of the (evolving) sets NBDFOUND and SETTLED. If a vertex $i \in [p]$ is such that $d_i = |N(i)| \leq \ell/2$, then NBDFOUND contains i, and similarly if every $j \in N(i)$ is such that $d_j \leq \ell/2$, then SETTLED contains i. Next, we observe that since we are conditioning on $\mathcal{E}_1, \ldots, \mathcal{E}_{k-1}$, the following statements hold provided $\ell \leq d_{\max}$:

- If i is such that $d_i \leq \ell$, then $\overline{N}(i) \cap \mathrm{SETTLED} = \emptyset$, since each $j \in N(i)$ has at least one neighbor (viz., i) has not been enrolled in NBDFOUND. Therefore, by (C1), $\widehat{N}(i)$, the output of `nbdSelect`$(i, \ell, \left\{X^{(j)}_{[p]\setminus\mathrm{SETTLED}}\right\}_{j \in S_1})$ is exactly $N(i)$ with probability at least $1 - \delta/2pd_{\max}$.
- For such an i (i.e., with $d_i \leq \ell$), `nbdSelect`$(i, \widehat{N}(i), \{X^{(j)}_{[p]\setminus\mathrm{SETTLED}}\}_{j \in S_2})$

returns **true**. On the other hand, if i is such that $d_i > \ell$, the subroutine $\texttt{nbdVerify}(i, \widehat{N}(i), \{X^{(j)}_{[p]\setminus\text{SETTLED}}\}_{j\in S_2})$ returns **false** since $\left|\widehat{N}(i)\right| \leq \ell$ by definition of the $\texttt{nbdSelect}$ function. Both these follow from (C2) and with probability at least $1-\delta/2pd_{\max}$.

Both these observations together imply that for any k such that $2^{k-1} \leq d_{\max}$, $\mathbb{P}\left[\mathcal{E}_k^c|\mathcal{E}_1,\ldots,\mathcal{E}_{k-1}\right] \leq p/d_{\max}$; observe that we have (quite conservatively) bounded this event by using p as an upper bound to the number of vertices whose degrees do not exceed 2^{k-1}. On the other hand, if $2^{k-1} > d_{\max}$, by observations made above, NBDFOUND $= [p]$. Therefore, Algorithm 1 would have terminated before k reached such a value and $\mathbb{P}\left[\mathcal{E}_k^c|\mathcal{E}_1,\ldots,\mathcal{E}_{k-1}\right] = 0$. Therefore, from (1), we have that the probability that the algorithm errs is no more than δ, as required.

Finally, observe that the above argument implies that with probability greater $1 - pd_{\max}\delta$, the following is true. Each vertex $i \in [p]$ is enrolled in NBDFOUND no later than when the counter reaches $\ell = 2^{\lceil \log_2 d_i \rceil} \leq 2d_i$. Therefore, by the time ℓ reaches $2d^i_{\max}$, every neighbor of i has already been enrolled in NBDFOUND, which of course implies that i is enrolled in SETTLED and is no longer sampled from. Therefore, the total number of samples accumulated for vertex i is given by $\sum_{k=0}^{\lceil \log_2 d^i_{\max}\rceil} g(2^k) + h(2^k)$. This implies that a budget $B \geq \sum_{i\in[p]} \sum_{0\leq k\leq \lceil \log_2 d^i_{\max}\rceil} g(2^k) + h(2^k)$ is sufficient. □

B Proof of Theorem 2

We will restate Theorem 2 here for convenience.

Theorem 2. *Fix $\delta \in (0,1)$ and suppose that assumptions (A1) and (A2) hold. Then, there exists a constant $c = c(m, M, \delta)$ such that if we set $g(\ell) = \lceil c\ell \log p\rceil$ and $\xi = m/2$, then with probability no less than $1-\delta$, the following hold:*

1. *The AdPaCT algorithm successfully recovers the graph G.*
2. *The computational complexity of the AdPaCT algorithm is no worse than $\mathcal{O}(p^{d_{\max}+2})$*

This implies that the total sample complexity of the AdPaCT algorithm at confidence level $1-\delta$ is bounded by $2c\bar{d}_{\max} p \log p$.

Proof. To prove Theorem 2, we will bound the probability that the conditions (C1) and (C2) are violated.

First, fix an arbitrary $\ell \leq d_{\max}$, a vertex $i \in [p]$, and a subset $F \subseteq [p]$ such that $d_i \leq \ell$ and $\overline{N}(i) \subseteq F$. For ease of notation, we will let $n_\ell = g(\ell)$. The event that the **nbdVerify** subroutine defined in Algorithm 2 does not satisfy the condition (C1) is equivalent to saying that there is a set $S \subseteq F$ such that $|S| \leq \ell$ and $\max_{j\notin S\cup\{i\}} \left|\widehat{\rho}_{i,j|S}\right| \leq \xi$ and one of the following events hold:

- $\max_{j\notin \overline{N}(i)} \left|\widehat{\rho}_{i,j|N(i)}\right| > \xi$
- $|S| < d_i$
- $|S| = d_i$ and $\max_{j\notin S\cup\{i\}} \left|\widehat{\rho}_{i,j|S}\right| \leq \max_{j\notin \overline{N}(i)} \left|\widehat{\rho}_{i,j|N(i)}\right|$.

Letting $\mathcal{S}_{i,j,\ell}$ denote the set of all sets of size at most ℓ that **do not** separate[1] i from j in the graph G, observe that above events imply that one or both of the following conditions hold: (a) there exists a vertex $j \in [p]\setminus\{i\}$ and a subset $S \subseteq \mathcal{S}_{i,j,\ell}$ such that $\left|\widehat{\rho}_{i,j|S}\right| \leq \xi$, or (b) there is a vertex $j \in [p] \setminus \overline{N}(i)$: $\left|\widehat{\rho}_{i,j|N(i)}\right| > \xi$. Therefore, we will bound the probability that (C1) does not hold, an event we will dub $\mathcal{E}_1$, as follows

$$\mathbb{P}\left[\mathcal{E}_1\right] \leq \sum_{\substack{j\in[p]\setminus\{i\} \\ S\in\mathcal{S}_{i,j,\ell}}} \mathbb{P}\left[\left|\widehat{\rho}_{i,j|S}\right| \leq \xi\right] + \sum_{j\in[p]\setminus\overline{N}(i)} \mathbb{P}\left[\left|\widehat{\rho}_{i,j|N(i)}\right| > \xi\right]. \tag{2}$$

To proceed, let us consider an arbitrary term in the first sum. Since $S \in \mathcal{S}_{i,j,\ell}$, we know by the second part of assumption (A2) that $\left|\rho_{i,j|S}\right| > m$. Now, observe that $\left|\widehat{\rho}_{i,j|S}\right| \leq \xi$ and $\left|\rho_{i,j|S}\right| \geq m$ together imply that $\left|\rho_{i,j|S}\right| - \left|\widehat{\rho}_{i,j|S}\right| \geq m - \xi \Rightarrow \left|\rho_{i,j|S} - \widehat{\rho}_{i,j|S}\right| \geq m - \xi$, since $m > \xi$. Therefore, in this case, we have

$$\begin{aligned}
&\mathbb{P}\left[\left|\widehat{\rho}_{i,j|S}\right| \leq \xi\right] \\
&\leq \mathbb{P}\left[\left|\widehat{\rho}_{i,j|S} - \rho_{i,j|S}\right| \geq m - \xi\right] && (3)\\
&\overset{(a)}{\leq} C_1\left(n_\ell - 2 - |S|\right) \\
&\qquad \exp\left\{-\left(n_\ell - 4 - |S|\right)\log\left(\frac{4+(m-\xi)^2}{4-(m-\xi)^2}\right)\right\} && (4)\\
&\leq C_1 n_\ell \exp\left\{-\left(n_\ell - 4 - \ell\right)\log\left(\frac{16+m^2}{16-m^2}\right)\right\}, && (5)
\end{aligned}$$

where (a) follows from Lemma 3 in Appendix D and as in the lemma, the constant C_1 only depends on M. The last step follows from the condition that $|S| \leq \ell$ and $\xi = m/2$.

[1] A set S is said to separate a pair of vertices i and j in a graph if all the paths that connect i and j contain at least one vertex from S.

Next, we will consider an arbitrary term in the second summation of (2). Since $j \notin \overline{N}(i)$, we know by assumption (A1), $\rho_{i,j|N(i)} = 0$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left[\left|\widehat{\rho}_{i,j|S}\right| \geq \xi\right] \\
&\quad = \mathbb{P}\left[\left|\widehat{\rho}_{i,j|S} - \rho_{i,j|S}\right| \geq \xi\right] && (6) \\
&\quad \overset{(a)}{\leq} C_1\left(n_\ell - 2 - |S|\right) \\
&\qquad \exp\left\{-\left(n_\ell - 4 - |S|\right)\log\left(\frac{4+\xi^2}{4-\xi^2}\right)\right\} && (7) \\
&\quad \leq C_1 n_\ell \exp\left\{-\left(n_\ell - 4 - \ell\right)\log\left(\frac{16+m^2}{16-m^2}\right)\right\}, && (8)
\end{aligned}$$

where, again, (a) follows from Lemma 3 in Appendix D and the last step follows after observing that $|S| \leq \ell$ and $\xi = m/2$. So, from (2), we have the following upper bound on the probability that (C1) is violated:

$$\begin{aligned}
&\mathbb{P}\left[\text{(C1) is violated for a fixed } i\right] \\
&\leq \sum_{\substack{j\in[p]\setminus\{i\} \\ S\in\mathcal{S}_{i,j,\ell}}} \mathbb{P}\left[\left|\widehat{\rho}_{i,j|S}\right| \leq \xi\right] + \sum_{j\in[p]\setminus\overline{N}(i)} \mathbb{P}\left[\left|\widehat{\rho}_{i,j|N(i)}\right| > \xi\right] && (9) \\
&\leq 2C_1 p^{\ell+1} n_\ell \exp\left\{-\left(n_\ell - 4 - \ell\right)\log\left(\frac{16+m^2}{16-m^2}\right)\right\}, && (10)
\end{aligned}$$

where the last step follows from observing that there are no more than $p^{\ell+1}$ terms in the first sum and p terms in the second sum. As observed in Section 4 in the manuscript, (C2) is satisfied by default for Algorithm 2 since the verification is implicitly performed by the exhaustive searching of the **ndbSelect** subroutine. Therefore, if the number of samples $n_\ell = g(\ell)$ satisfies

$$n_\ell \geq \left\lceil \ell + 5 + \frac{\left((\ell+5)\log(2C_1 p) + \log(2/\delta)\right)}{\log\left(\frac{16+m^2}{16-m^2}\right)} \right\rceil,$$

then the following implications hold

$$\begin{aligned}
&2C_1 p^{\ell+1} n_\ell \exp\left\{-\left(n_\ell - 4 - \ell\right)\log\left(\frac{16+m^2}{16-m^2}\right)\right\} \\
&\leq 2C_1 p^{\ell+1}\left[\ell + 6 + \frac{\left((\ell+5)\log(2C_1 p) + \log(2/\delta)\right)}{\log\left(\frac{16+m^2}{16-m^2}\right)}\right] \\
&\qquad \exp\left\{-\left(\ell+5\right)\log(2C_1 p)\right\} \times \frac{\delta}{2} && (11) \\
&\leq (2C_1)^{-\ell-4} p^{-4} \\
&\qquad \left[p + 6 + \frac{\left((p+5)\log(2C_1 p) + \log(2/\delta)\right)}{\log\left(\frac{16+m^2}{16-m^2}\right)}\right] && (12) \\
&\leq \frac{\delta}{2} p^{-2}, \text{ for } p \text{ large enough.} && (13)
\end{aligned}$$

Therefore, from Theorem 1, we have that the AdPaCT algorithm succeeds with probability exceeding $1 - pd_{\max}\delta p^{-2} \geq 1 - \delta$ (since $d_{\max} \leq p$), as required. The above calculation also demonstrates that (provided there is a constant $c'' > 0$ such that $\delta < p^{-c''}$, which is a mild requirement) there exists a constant $c' > 0$ such that we might choose $g(\ell) = \lceil c'\ell \log(p)/2 \rceil$. Finally, using Theorem 1, we observe that the number of samples accumulated for vertex i is no more than $\sum_{k \leq \lceil \log_2 d^i_{\max} \rceil} g(2^k) + h(2^k) \leq 2cd^i_{\max} \log p/2$ (where c accounts for the integer effects). Therefore, for the AdPaCT algorithm to succeed, it suffices to pick a budget that satisfies $B \geq c\bar{d}_{\max} p \log p$. Finally, observe that the computational complexity statement follows from the size of the subsets that are being searched over. This concludes the proof. □

C Proof of Theorem 3

In this section, we will prove Theorem 3, which we again restate here for convenience.

Theorem 3. *Fix $\delta > 0$. Suppose that assumptions (A1)-(A4) hold. There exists constants C_1, C_2, C_3 which depend on Σ, m, δ such that if we set $c = C_1$ (i.e., $g(\ell) = c\ell \log p$), $\xi = m/2$, $\lambda_\ell = \sqrt{\frac{2C_2\|\Sigma\|_\infty}{C_1\gamma^2}}$, and budget $B = 2c\bar{d}_{\max} p \log p$, then with probability at least $1 - \delta$, the following hold*

1. *the AMPL algorithm successfully recovers the graph G,*
2. *The computational complexity is bounded from above by $d_{\max} p \mathfrak{C}$, where $\mathfrak{C}$ is the computational cost of solving a single instance of Lasso,*

provided

$$\begin{aligned}
m \geq &\left(\frac{C_{\min}}{C_{\max}} + \frac{C_{\max}}{C_{\min}} + 2\right) \times \\
&\frac{1}{4\min_i |\Sigma_{ii}|}\left[C_3\sqrt{\frac{2C_1\|\Sigma\|_\infty}{C_2\gamma^2}} \max_i \left|\left|\left|\left(\widetilde{\Sigma}^i_{N(i),N(i)}\right)^{-1/2}\right|\right|\right|^2_\infty \right. \\
&\qquad \left. + 20\sqrt{\frac{\|\Sigma\|_\infty}{C_{\min}C_2}}\right]
\end{aligned}$$

Proof. As in the proof of Theorem 2, we will prove Theorem 3 by showing that the subroutines `nbdSelect` and `nbdVerify` from Algorithm 3 (AMPL) satisfy the conditions (C1) and (C2) of Theorem 1. Along the way, we will also identify the functions $g()$ and $h()$ which will suggest a choice for the budget.

Lemma 1. *Fix an arbitrary $\ell \leq d_{\max}$. Let $i \in [p]$ and $F \subseteq [p]$ be such that $d_i \leq \ell$ and $\overline{N}(i) \subseteq F$. There exist constants $C_1, C_2, C_3 > 0$ such that the* **ndbSelect**

subroutine returns $N(i)$ with probability greater than $1-\delta/pd_{\max}$, provided the following hold

1. $g(\ell) = C_1 \ell \log p$
2. $\lambda_\ell = \sqrt{\frac{2C_2\|\Sigma\|_\infty}{C_1\gamma^2}}$
3. $$m \geq \left(\frac{C_{\min}}{C_{\max}} + \frac{C_{\max}}{C_{\min}} + 2\right) \times \frac{1}{4\min_i |\Sigma_{ii}|} \times \left[C_3\sqrt{\frac{2C_1\|\Sigma\|_\infty}{C_2\gamma^2}} \max_i \left|\left|\left|\left(\widetilde{\Sigma}^i_{N(i),N(i)}\right)^{-1/2}\right|\right|\right|_\infty^2 + 20\sqrt{\frac{\|\Sigma\|_\infty}{C_{\min}C_2}}\right]$$

Proof. The subroutine **ndbSelect** receives as input $g(\ell)$ (which we will denote as n_ℓ for the rest of this proof) samples from the random variables $X_F = \{X_i : i \in F\}$. Notice that X_F is distributed according to $\mathcal{N}(0, \bar{\Sigma})$, where we write $\bar{\Sigma}$ to denote $\Sigma(F,F)$, the (F,F) submatrix of Σ.

First, we will begin by providing justification for the fact that the Lasso [1] can be used for selecting the neighborhood of i in our setting. The seminal work of [2] first recognized that the Lasso can be used for neighborhood selection in Gaussian graphical models. We will adapt this insight, while accounting for the sequential marginalization of our active algorithm. Towards this end, let X_i denote the random variable corresponding to vertex i and let X_G denote the random vector corresponding to the vertices $G \triangleq F \setminus \{i\}$. As noted above, $X_F \sim \mathcal{N}(0, \bar{\Sigma})$. Notice that the corresponding precision matrix, $\bar{K} \triangleq \bar{\Sigma}^{-1}$, is not equal to the original precision matrix K. We know that conditioned on X_G, X_i behaves like a Gaussian random variable, and in particular, the conditional distribution takes the following form (see, e.g., [3, Chapter 9]):

$$p(X_i \mid X_G) = \mathcal{N}(\bar{\Sigma}_{iG}\bar{\Sigma}_{GG}^{-1}X_G, \bar{\Sigma}_{ii} - \bar{\Sigma}_{iG}\bar{\Sigma}_{GG}^{-1}\bar{\Sigma}_{Gi}). \tag{14}$$

Now, we will let $y \in \mathbb{R}^{n_\ell}$ denote the vector of samples from X_i that is received by **ndbSelect** and we will let $\mathbf{X} \in \mathbb{R}^{n_\ell \times |G|}$ denote the matrix of corresponding samples from the random vector X_G. This notation will let us simplify the following presentation while allowing us to readily identify the components of our problem with the classic Lasso problem. With this notation, from (14), we can write down the "true model" corresponding to our problem as

$$y = \mathbf{X}\beta^* + w, \tag{15}$$

where $y, \mathbf{X}$ are as above, $\beta^* \triangleq \bar{\Sigma}_{iG}\bar{\Sigma}_{GG}^{-1} \in \mathbb{R}^{|G|}$, and $w \in \mathbb{R}^{n_\ell} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \bar{\Sigma}_{ii} - \bar{\Sigma}_{iG}\bar{\Sigma}_{GG}^{-1}\bar{\Sigma}_{Gi}\right)$. We will now show that recovering the support of β^* suffices.

Claim 1. *For any $j \in F \setminus \{i\}$, we have that $\beta^*_j = -\frac{K_{ij}}{K_{ii}}$.*

Proof. To prove this, we will first begin by writing $\bar{\Sigma}$ and $\bar{K}$ in their partitioned form:

$$\bar{\Sigma} = \begin{pmatrix} \bar{\Sigma}_{ii} & \bar{\Sigma}_{iG} \\ \bar{\Sigma}_{Gi} & \bar{\Sigma}_{GG} \end{pmatrix} \tag{16}$$

$$\bar{K} = \begin{pmatrix} \bar{K}_{ii} & \bar{K}_{iG} \\ \bar{K}_{Gi} & \bar{K}_{GG} \end{pmatrix} \tag{17}$$

Using a standard block matrix inversion (see, e.g., [4]), we observe that

$$\begin{aligned} \begin{pmatrix} \bar{K}_{ii} & \bar{K}_{iG} \\ \bar{K}_{Gi} & \bar{K}_{GG} \end{pmatrix} &= \bar{K} = \bar{\Sigma}^{-1} \\ &= \begin{pmatrix} \bar{\Sigma}_{ii} & \bar{\Sigma}_{iG} \\ \bar{\Sigma}_{Gi} & \bar{\Sigma}_{GG} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \bar{\Sigma}_{i\backslash G}^{-1} & -\bar{\Sigma}_{i\backslash G}^{-1}\bar{\Sigma}_{iG}\bar{\Sigma}_{GG}^{-1} \\ * & * \end{pmatrix}, \end{aligned} \tag{18}$$

where $\bar{\Sigma}_{i\backslash G} = \bar{\Sigma}_{ii} - \bar{\Sigma}_{iG}\bar{\Sigma}_{GG}^{-1}\bar{\Sigma}_{Gi}$, and the $*$ values maybe ignored for the present calculation. Comparing the two block matrices above, it becomes clear that $-\bar{K}_{ii}\bar{\Sigma}_{iG}\bar{\Sigma}_{GG}^{-1} = \bar{K}_{iG}$. Since we know from (14) that $\beta^* = \bar{\Sigma}_{iG}\bar{\Sigma}_{GG}^{-1}$, we have shown that $\beta^*_j = -\frac{\bar{K}_{ij}}{\bar{K}_{ii}}$.

Recall that $\bar{\Sigma} = \Sigma(F,F)$. Therefore, arguing as above, we have the following relationship between the entries of K and $\bar{K}$

$$\bar{K} = K_{FF} - K_{FF^c}K_{F^cF^c}^{-1}K_{F^cF}. \tag{19}$$

$\bar{K}$ is called the *Schur complement* of the block $K_{F^cF^c}$ with respect to the matrix K. Now, by the hypothesis of the condition (C1), we know that $F^c \subseteq \bar{N}(i)^c$. This implies that $K_{iF^c} = 0$ identically, and in particular, this means that $\bar{K}_{ij} = K_{ij}$ for any $j \in F$. This concludes the proof. □

Along with the fact that the non-zeros in the concentration matrix, K, correspond to graph edges, Claim 1 shows us that the support of β^* from (15) gives us the neighborhood of i.

Observe that the candidate neighborhood chosen by the **ndbSelect** subroutine is the support of a vector $\widehat{\beta} \in \mathbb{R}^{|G|}$, where

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^{|G|}}{\arg\min} \frac{1}{2n_\ell}\|y - \mathbf{X}\beta\|_2^2 + \lambda_\ell \|\beta\|_1. \tag{20}$$

Therefore, to conclude the proof of Lemma 1, it suffices to show that $\widehat{\beta}$ and β^* have the same support with high probability. For this, we will borrow the results of Theorem 4, which is based on the seminal work of Wainwright [5]. In order to apply this theorem in our setting, we need to ensure that the assumptions of the theorem are satisfied, and this is what the next claim will demonstrate.

Claim 2. *Let $\breve{\Sigma} \triangleq \Sigma(F \setminus \{i\}, F \setminus \{i\})$ denote the covariance matrix corresponding to the rows of $\mathbf{X}$ and let $\beta^*_{\min} = \min_{j \in N(i)} |\beta^*_j|$. Then, the following hold*

$$\left\|\left\|\breve{\Sigma}_{F\setminus\overline{N}(i),N(i)}\left(\breve{\Sigma}_{N(i),N(i)}\right)^{-1}\right\|\right\|_{\infty} \leq 1-\gamma \tag{21}$$

$$0 < C_{\min} \leq \Lambda_{\min}\left(\breve{\Sigma}_{N(i),N(i)}\right) \tag{22}$$

$$\Lambda_{\max}\left(\breve{\Sigma}_{N(i),N(i)}\right) \leq C_{\max} < \infty. \tag{23}$$

$$\beta^*_{\min} \geq \frac{m \times \min_{i\in[p]} \Sigma_{ii}}{\left(\frac{C_{\min}}{C_{\max}} + \frac{C_{\max}}{C_{\min}} + 2\right)} \tag{24}$$

Proof. First observe that the submatrices $\breve{\Sigma}_{N(i),N(i)}$ and $\widetilde{\Sigma}_{N(i),N(i)}$ are identical since, by assumption $N(i) \subseteq F \setminus \{i\}$. Therefore, by the hypotheses of Theorem 3, the second set of inequalities follow immediately. Similarly, we observe that since $F \setminus \overline{N}(i) \subseteq [p] \setminus \overline{N}(i)$, the hypothesis of Theorem 3 implies that

$$\begin{aligned}
&\left\|\left\|\breve{\Sigma}_{F\setminus\overline{N}(i),N(i)}\left(\breve{\Sigma}_{N(i),N(i)}\right)^{-1}\right\|\right\|_{\infty}\\
&= \max_{j\in F\setminus\overline{N}(i)} \sum_{r\in N(i)} \left|\sum_{t\in\overline{N}(i)} \breve{\Sigma}_{j,r}\left[\left(\breve{\Sigma}_{N(i),N(i)}\right)^{-1}\right]_{r,t}\right|\\
&= \max_{j\in F\setminus\overline{N}(i)} \sum_{r\in N(i)} \left|\sum_{t\in\overline{N}(i)} \breve{\Sigma}_{j,r}\left[\left(\widetilde{\Sigma}_{N(i),N(i)}\right)^{-1}\right]_{r,t}\right|\\
&\leq \max_{j\in [p]\setminus\overline{N}(i)} \sum_{r\in N(i)} \left|\sum_{t\in\overline{N}(i)} \breve{\Sigma}_{j,r}\left[\left(\widetilde{\Sigma}_{N(i),N(i)}\right)^{-1}\right]_{r,t}\right|\\
&\leq 1-\gamma.
\end{aligned} \tag{25}$$

To conclude the proof of the claim, we will provide a lower bound on $\beta^*_{\min}$. Towards this end, note that by Claim 1, we have that

$$\beta^*_{\min} \geq \min_{i,j\in[p]:K_{ij}\neq 0} \frac{|K_{ij}|}{\max\{|K_{ii}|\,.\,|K_{jj}|\}} \geq \frac{m}{\max_{i\in[p]}|K_{ii}|}, \tag{26}$$

where the last inequality follows from assumption (A2). We will proceed by obtaining an upper bound on the denominator. Towards this end, we will employ the well known Kantorovich inequality for positive definite matrices (see, e.g., [6]). This inequality states that for a positive definite matrix $A \in \mathbb{R}^{d\times d}$ with real eigenvalues $L \leq \lambda_1 \leq \cdots \leq \lambda_d \leq U$, the following holds for any $x \in \mathbb{R}^d$

$$1 \leq \left(x^T A x\right)\left(x^T A^{-1} x\right) \leq \frac{1}{4}\left(\frac{L}{U} + \frac{U}{L} + 2\right). \tag{27}$$

Therefore, choosing x to be the $i-$th canonical vector e_i, we have the following useful inequality relating the diagonal elements of a matrix to the diagonal elements of its inverse

$$A^{-1}_{ii} \leq \frac{1}{4A_{ii}}\left(\frac{L}{U} + \frac{U}{L} + 2\right). \tag{28}$$

Applying this to (26), we get the desired result. This concludes the proof of the claim. □

Claim 2 paves the way for applying Theorem 4 to our setting. Before, we conclude, we observe the following:

(a) The noise variance (σ^2) of Theorem 4 can be taken to be $\max_{i\in[p]} \Sigma_{ii}$, from (15). It is not hard to see that since Σ is a positive definite matrix, we have that $\max_{i\in[p]} \Sigma_{ii} = \|\Sigma\|_\infty$, the absolute maximum element of Σ.

(b) Our bound from Claim 2 on $\beta^*_{\min}$ implies that we can satisfy the so-called "beta-min" condition required by Theorem 4, provided m is as in the statement of Theorem 3

Therefore, we now have that there exists constants $C_1, C_2 > 0$ such that if we set $n_\ell = C_1 \ell \log p$ and $\lambda_\ell = \sqrt{\frac{2C_2\|\Sigma\|_\infty}{C_1\ell\gamma^2}}$, we can be guaranteed that **ndbSelect** succeeds with probability at least $1 - \delta/2pd_{\max}$.

This concludes the proof. □

Lemma 2. *Fix an arbitrary $\ell \leq d_{\max}$, a vertex $i \in [p]$ and subsets $F, G \subseteq [p]$ that are such that $d_i \leq \ell$, $\overline{N}(i) \subseteq F$, and $H \subseteq F$. There exists a constant $C_4 > 0$ such that if we set $h(\ell) = C_4 \ell \log p$, then the probability that* **nbdVerify** *fails (as in (C2)) is at most $\delta/2pd_{\max}$.*

The proof of this lemma follows directly from the proof of Theorem 2. Therefore, using Lemma 1 and Lemma 2 in Theorem 1, completes the proof of Theorem 3. □

D Helpful Results

D.1 Concentration of Partial Correlation Coefficients

In this section, we will state a lemma that characterizes the concentration of empirical partial correlation coefficients (defined as in the paragraph after equation (3) in the manuscript) about their expected values. See [7] for a proof.

Lemma 3. *Provided (A2) holds, given n samples from (X_i, X_j, X_S), if the partial correlation coefficient $\widehat{\rho}_{i,j|S}$ is defined as above, then we have the following result*

$$\mathbb{P}\left[\left|\widehat{\rho}_{i,j|S} - \rho_{i,j|S}\right| \geq \epsilon\right] \leq C_1 \left(n - 2 - |S|\right) exp\left\{-\left(n - 4 - |S|\right)\log\left(\frac{4+\epsilon^2}{4-\epsilon^2}\right)\right\}, \tag{29}$$

where $C_1 > 0$ is a constant that depends on M from (A2).

D.2 Support Recovery for Lasso

Assume that $y = X\beta^* + w$, where $y, w \in \mathbb{R}^n$, $\beta^* \in \mathbb{R}^p$, and $X \in \mathbb{R}^{n\times p}$ with iid rows $x_i \sim \mathcal{N}(0, \Sigma)$. Suppose S is the support of β^* and suppose that the following hold

$$\left|\left|\left|\Sigma_{S^cS}\left(\Sigma_{SS}\right)^{-1}\right|\right|\right|_\infty \leq 1-\gamma, \gamma \in (0,1] \tag{30}$$

$$\Lambda_{\min}\left(\Sigma_{SS}\right) \geq C_{\min} > 0 \tag{31}$$

$$\Lambda_{\max}\left(\Sigma_{SS}\right) \leq C_{\min} < +\infty \tag{32}$$

If we let $\widehat{\beta} \in \mathbb{R}^p$ denote the solution to the Lasso problem

$$\widehat{\beta} \triangleq \frac{1}{2n}\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \tag{33}$$

then we have the following result.

Theorem 4. *Suppose $w \sim \mathcal{N}(0, \sigma^2 I)$ and suppose that Σ satisfies the properties listed above. Then, there exists constants C_1, C_2, C_3, C_4, C_5 such that if $\lambda_n = \sigma\gamma^{-1}\sqrt{2C_1 \log p/n}$, $n \geq C_2 k \log p$, and $\beta_{\min} \triangleq \min_{i\in S}|\beta_i^*| > u(\lambda_n)$, where*

$$u(\lambda_n) \triangleq C_5\lambda_n \left|\left|\left|\Sigma_{SS}^{-1/2}\right|\right|\right|_\infty^2 + 20\sqrt{\frac{\sigma^2 \log k}{C_{\min} n}}, \tag{34}$$

the support $\widehat{\beta}$ is identical to that of β^* with probability exceeding $1 - C_3 p^{-C_4}$.

Proof. The proof of this theorem follows almost entirely from Theorem 3 in [5]. In fact, the only thing we modify from that result is the rate of decay of the probability of error. In particular, we will show that the probability of error decays polynomially in p (or equivalently exponentially in $\log p$) for all values of k, whereas Theorem 3 of [5] shows that the error decays exponentially in $\min\{k, \log(p-k)\}$, which is somewhat weak for our purposes.

Towards this end, it is not hard to see that the result that requires strengthening is Lemma 5 in [5]. We furnish a sharper substitute in Lemma 4. □

Lemma 4. *Consider a fixed $z \in \mathbb{R}^k$, a constant $c_1 > 0$, and a random matrix $W \in \mathbb{R}^{n\times k}$ with i.i.d elements $W_{ij} \sim \mathcal{N}(0,1)$. Suppose that $n \geq \max\left\{\frac{4}{(\sqrt{8}-1)^2}k, \frac{64}{c_1^2}k\log(p-k)\right\}$, then there exists a constant $c_2 > 0$ such that*

$$\mathbb{P}\left[\left\|\left[\left(\frac{1}{n}W^TW\right)^{-1} - I_k\right]z\right\|_\infty \geq C_1\|z\|_\infty\right] \leq 4\exp\left(-c_2\log(p-k)\right)$$

Proof. Set $A = \left(\frac{1}{n}W^TW\right)^{-1} - I_k$. Observe that $\mathbb{P}\left[\|Az\|_\infty \geq c_1\|z\|_\infty\right] \leq \mathbb{P}\left[\|A\|_\infty \geq c_1\right]$ by the definition of the matrix infinity norm. Next, observe that since the infinity norm is the maximum absolute row sum of the matrix, we have that $\mathbb{P}\left[\|A\|_\infty \geq c_1\right] \leq \mathbb{P}\left[\|A\|_2 \geq c_1/\sqrt{k}\right]$. From [5, Lemma 9] (which follows in a straightforward manner from the seminal results of [8]), we know that

$$\mathbb{P}\left[\|A\|_2 \geq \delta(n,k,t)\right] \leq 2e^{-nt^2/2}, \tag{35}$$

where $\delta(n,k,t) = 2\left(\sqrt{\frac{k}{n}} + t\right) + \left(\sqrt{\frac{k}{n}} + t\right)^2$. We will divide the proof into three cases:

Case (a): $k \leq \frac{c_1^2}{64}$

Suppose we pick $t = \sqrt{\frac{c_1}{\sqrt{k}}} - 1 - \sqrt{\frac{k}{n}}$, under the setting of this case, provided $n \geq \frac{4}{(\sqrt{8}-1)^2}k$, we have that $t > \frac{\sqrt{8}-1}{2} > 0$. Notice that for this choice of t, we have $\delta(n,k,t) = \frac{c_1}{\sqrt{k}}$. This gives us the following bound

$$\mathbb{P}\left[\|A\|_2 \geq \frac{c_1}{\sqrt{k}}\right] \leq 2\exp\left\{-\frac{n}{2}\left(\sqrt{\frac{c_1}{\sqrt{k}}} - 1 - \sqrt{\frac{k}{n}}\right)^2\right\} \tag{36}$$

$$\leq 2\exp\left\{-\frac{n}{2}\left(\frac{\sqrt{8}-1}{2}\right)^2\right\} \tag{37}$$

Case (b): $\log(p-k) \geq k > \frac{c_1^2}{64}$

Suppose we pick $t = \frac{c_1}{8\sqrt{k}}$, we have that $t < 1$, by the assumption of this case. Then, if $n \geq \frac{64k^2}{c_1^2}$ observe that

$$\delta(n,k,t) = 2\left(\sqrt{\frac{k}{n}} + t\right) + \left(\sqrt{\frac{k}{n}} + t\right)^2 \tag{38}$$

$$\leq \frac{c_1}{\sqrt{k}}. \tag{39}$$

This implies that

$$\mathbb{P}\left[\|A\|_2 \geq \frac{c_1}{\sqrt{k}}\right] \leq 2\exp\left\{-\frac{nc_1^2}{128k}\right\}. \tag{40}$$

Notice that if $n \geq \frac{64}{c_1^2} k \log(p-k)$, then $n \geq \frac{64k^2}{c_1^2}$, as required.

Case (c): $k > \log(p-k)$

In this case, we can adopt the result from Lemma 5 of [5].

Putting all this together, we get the desired result. □

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[2] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, pp. 1436–1462, 2006.

[3] J. A. Gubner, *Probability and random processes for electrical and computer engineers.* Cambridge University Press, 2006.

[4] G. H. Golub and C. F. Van Loan, *Matrix computations*, vol. 3. JHU Press, 2012.

[5] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso)," *Information Theory, IEEE Transactions on*, vol. 55, no. 5, pp. 2183–2202, 2009.

[6] R. A. Horn and C. R. Johnson, *Matrix analysis.* Cambridge university press, 2012.

[7] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the pc-algorithm," *The Journal of Machine Learning Research*, vol. 8, pp. 613–636, 2007.

[8] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices and banach spaces," *Handbook of the geometry of Banach spaces*, vol. 1, pp. 317–366, 2001.