Active Learning Algorithms for Graphical Model Selection

Anonymous Authors
Unknown Institution

A Proof of Theorem 1

We will let $\mathcal{E}_\ell$ denote the event that at step ℓ, the following holds: for each $i \in [p] \backslash \textsc{nbdFound}$, (a) If $|N(i)| \leq \ell$, then $\widehat{N}(i)$, the output of $\texttt{nbdSelect}\left(i, \ell, \{X^{(j)}_{[p]\backslash \textsc{settled}}\}_{j \in S_1}\right)$ is identical to $N(i)$, and furthermore, $\texttt{nbdVerify}(i, \widehat{N}(i), \{X_{[p]\backslash \textsc{settled}}\}_{j \in S_1}) = \textbf{true}$.

(b) If $|N(i)| > \ell$, then $\texttt{nbdVerify}(i, \widehat{N}(i), \{X_{[p]\backslash \textsc{settled}}\}_{j \in S_2})$ returns **false**.

Let us condition on $\bigcap_\ell \mathcal{E}_\ell$. Observe that if $i \notin \textsc{nbdFound}$ at step ℓ, then we are guaranteed that $\bar{N}(i) \subset [p] \backslash \textsc{settled}$. Therefore, if $d_i \leq \ell$, we are guaranteed that `nbdSelect` will correctly identify $N(i)$ and `nbdVerify` will confirm this identification. On the other hand, if $d_i > \ell$, we are guaranteed that $N(i) \not\subset \widehat{N}(i)$ (since $|\widehat{N}(i)| = d_i \leq \ell$). Therefore, `nbdVerify` will return **false**. These two properties together imply that Algorithm 1 correctly learns the graph. Additionally observe that when counter $\ell = d_i$, vertex i is enrolled in $\textsc{nbdFound}$ and by the time the counter ℓ reaches $d^i_{\max}$, every neighbor of i has already been enrolled in $\textsc{nbdFound}$, which of course implies that i is enrolled in $\textsc{settled}$ and is no longer sampled from. Therefore, the total number of samples accumulated for vertex i is given by $g(d^i_{\max}) + h(d^i_{\max})$. This implies that a budget $B \geq \sum_{i \in [p]} g(d^i_{\max}) + h(d^i_{\max})$ is sufficient.

To conclude the proof, we simply observe from (C1) and (C2) (and a union bound) that $\mathbb{P}[\mathcal{E}^c_\ell] \leq p\delta$, and again using the union bound over the levels ℓ, we get the desired result.

B Proof of Theorem 2

In order to prove Theorem 2, we will let $\mathcal{E}$ denote the event that all the conditional independence tests succeed. That is for a pair of vertices $i, j \in [p]$ and a subset $S \subset [p] \setminus \{i, j\}$, if $X_i \perp\!\!\!\perp X_j \mid X_S$, then $\left|\widehat{\rho}_{i,j|S}\right| \leq \xi$ and alternatively if $X_i \not\perp\!\!\!\perp X_j \mid X_S$, then $\left|\widehat{\rho}_{i,j|S}\right| > \xi$. Conditioned on $\mathcal{E}$, it is clear that the following hold

1. Since (A1) holds, `nbdFound`, as defined in Algorithm 2, satisfies condition (C1) from the statement of Theorem 1.
2. `nbdVerify`, as defined in Algorithm 2, satisfies condition (C2) from the statement of Theorem 1.
3. Therefore, from Theorem 1, we know that Algorithm 2 succesfully recovers the graph.
4. Also, Algorithm 2 terminates when ℓ reaches $d_{\max}$. This implies that the computational complexity of Algorithm 2 scales as $\mathcal{O}(p^{d_{\max}+2})$.

Next, we will bound the probability that for a fixed ℓ `nbdFound` fails to satisfy condition (C1) of Theorem 1. Observe that, by the union bound, this probability is bounded from above by

$$\sum_{\substack{i,j \in [p], \\ S \subset [p]\backslash\{i,j\} : |S| = \ell}} \mathbb{P}\left[\mathcal{E}^c_{i,j|S}\right]. \tag{1}$$

We will now turn our attention to one of the inner terms. Let us suppose that we are at level ℓ and let $n(i, j, S)$ be the total number of samples collected of the random vector (X_i, X_j, X_S); we will just refer to this quantity as n in what follows. We can split up the analysis into two parts:

Case (A): $X_i \perp\!\!\!\perp X_j \mid X_S$: In this case, we have that $\rho_{i,j|S} = 0$. Therefore,

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

where (a) follows from Lemma 1 in Appendix D and as in the lemma, the constant C_1 only depends on M.

The last step follows from plugging in $|S| = \ell$ and $\xi = m/2$.

Case (B): $X_i \not\perp\!\!\!\perp X_j \mid X_S$: In this case, we have that $|\rho_{i,j|S}| > m$, by assumption (A2). Also observe that $|\widehat{\rho}_{i,j|S}| \leq \xi$ and $|\rho_{i,j|S}| \geq m$ together imply that $|\rho_{i,j|S}| - |\widehat{\rho}_{i,j|S}| \geq m - \xi \Rightarrow |\rho_{i,j|S} - \widehat{\rho}_{i,j|S}| \geq m - \xi$, since $m > \xi$. Therefore, in this case, we have

$$\begin{aligned}
&\mathbb{P}\left[\mathcal{E}^c_{i,j|S}\right] \\
&= \mathbb{P}\left[|\widehat{\rho}_{i,j|S}| \leq \xi\right] && (6)\\
&\leq \mathbb{P}\left[|\widehat{\rho}_{i,j|S} - \rho_{i,j|S}| \geq m - \xi\right] && (7)\\
&\overset{(b)}{\leq} C_1 (n-2-|S|) \\
&\quad \exp\left\{-(n-4-|S|)\log\left(\frac{4+(m-\xi)^2}{4-(m-\xi)^2}\right)\right\} && (8)\\
&= C_1 (n-2-\ell) \\
&\quad \exp\left\{-(n-4-\ell)\log\left(\frac{16+m^2}{16-m^2}\right)\right\}, && (9)
\end{aligned}$$

where again (b) follows from Lemma 1 in Appendix D and the last step follows from substituting $|S| = \ell$ and $\xi = m/2$.

Using (5) and (9) in (1), we see that the probability that `nbdFound` fails to satisfy condition (C1) of Theorem 1 for a fixed ℓ is bounded from above by

$$\sum_{\substack{i,j\in[p],\\ S\subset[p]\setminus\{i,j\}:|S|=\ell}} 2C_1 (n-2-\ell) \times \exp\left\{-(n-4-\ell)\log\left(\frac{16+m^2}{16-m^2}\right)\right\}.$$

Therefore, for all sufficiently large p [1], we can check that if n satisfies

$$n \geq \frac{2(\ell+3+C_2)}{\log\left(\frac{16+m^2}{16-m^2}\right)}\log\left(\frac{p}{(2C_1)^{C_2}}\right) + 4 + \ell, \qquad (10)$$

the above sum can be bounded from above by p^{-C_2-1} (note that this quantity plays the role of $p\delta$ in Theorem 1). Since, for $\ell \geq 1$, $2\ell(C_2+4) \geq 2(\ell+3+C_2)$, we can set $g(\ell) = c\log p$ with $c = \frac{4(C_2+4)}{\log\left(\frac{16+m^2}{16-m^2}\right)}$ in Algorithm 2. This would imply that for all sufficiently large p, [2] we have $n(i,j,S) \geq \ell c \log p$ which guarantees that (10) is satisfied for all ℓ. Now, since ℓ is never more than p, we have

$$\mathbb{P}[\text{error}] \leq p^{-C_2}. \qquad (11)$$

[1] n needs to be such that $(n-4-\ell) \geq \frac{2\log(n-2-\ell)}{\log((16+m^2)/(16-m^2))}$

[2] $p > (2C_1)^{-C_2} \wedge 4e/\ell$, where the latter part guarantees that the additive term $4+\ell$ is accounted for.

Since we just demonstrated that we can choose $g(\ell) = \ell c \log p$ (and take $h(\ell)$ to be 0), from Theorem 1 we know that it suffices to set the budget as in Theorem 2, i.e., $B \geq c\bar{d}_{\max} p \log p$. This concludes the proof.

C Proof of Theorem 3

As in the proof of Theorem 2, we will prove Theorem 3 by showing that the choice of `nbdSelect` and `nbdVerify` from Algorithm 3 satisfies the conditions 1 and 2 of Theorem 1. Along the way, we will also identify the functions $g()$, $h()$, and the probability of making an error in an intermediate step, δ. This will complete the proof.

Let us suppose that we are at iteration number ℓ and let $\mathcal{E}^c_\ell$ be the event that this is the first iteration where either `nbdSelect` or `nbdVerify` make an error. Notice that $\sum_{\ell\in[p]} \mathbb{P}[\mathcal{E}^c_\ell]$ bounds the probability of error from above.

`nbdSelect` satisfies condition (C1) of Theorem 1 Suppose that $d_i = \ell$. Since assumptions (A3) - (A4) hold[3], we know from Theorem 4, that there exists constants C_3, C_4, C_5 such that provided $g(\ell) = C_3 \ell \log p$, we are guaranteed that the probability that $\widehat{N}(i) \neq N(i)$ is bounded from above by $C_4 p^{-C_5}$. Therefore, the probability that there exists an i such that $|N(i)| = \ell$ whose neighborhood is not found is bounded from above by $C_4 p^{-(C_5-1)}$.

It is important to observe here that the regression coefficients of the Lasso problem remain unchanged even though we are observing marginal samples. To see why this is true, simply observe that the inverse covariance matrix corresponding to the variables in S (i.e., S^c is marginalized out) is given by $K_{SS} - K_{SS^c}\left(K_{S^cS^c}\right)^{-1} K_{S^cS}$, the appropriate Schur complement. Now conditioned on making no errors in the previous stages, we can see that the corresponding row of the precision matrix remains unchanged.

`nbdVerify` satisfies condition (C2) of Theorem 1 Suppose $d_i = \ell$, `nbdVerify` fails to satisfy condition (C2) of Theorem 1 implies that there is a failed conditional independence test. At iteration number ℓ, the probability of failure of a conditional independence test with a conditioning set of size ℓ is upper bounded as in the proof of Theorem 2 by p^{-C_2} provided we choose $h(\ell) = \frac{4(C_2+4)}{\log((16+m^2)/(16-m^2))}\ell \log p$. On the other hand, consider the situation that $d_i > \ell$. Since `nbdSelect`, as defined in Algorithm 3, truncates $\widehat{N}(i)$ to be of size ℓ, we are guaranteed that there is a $j \in N(i) \cap \widehat{N}(i)^c$. We can again bound this

[3] note that these assumptions continue to hold even after marginalizing out the variables in SETTLED.

by bounding the probability of a failed conditional independence test, which is the same as before.

We can now conclude the proof by observing that the choice for $g(\ell)$ and $h(\ell)$ as stated in the theorem allows one to bound the probability of error (and implicitly δ) from above by $\sum_{\ell=1}^{p} C_4 p^{-C_5+1} + p^{-C_2} + p^{-C_2}$. Choosing the constants appropriately, we conclude that the total sample complexity (and hence a valid choice for B) is given by $c\bar{d}_{\max} p \log p$ at a confidence level $1 - p^{-C_2}$.

D Helpful Results

D.1 Concentration of Partial Correlation Coefficients

In this section, we will record some useful lemmata. The first lemma concerns the concentration of empirical partial correlation coefficients (defined as in the paragraph after (2) in the manuscript) about their expected values. See [1] for a proof.

Lemma 1. *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*

$$\mathbb{P}\left[\left|\widehat{\rho}_{i,j|S} - \rho_{i,j|S}\right| \geq \epsilon\right] \leq C_1 (n - 2 - |S|) exp\left\{-(n - 4 - |S|) \log\left(\frac{4+\epsilon^2}{4-\epsilon^2}\right)\right\}, \tag{12}$$

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

D.2 Support Recovery for Lasso

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

$$\left\|\Sigma_{S^cS} (\Sigma_{SS})^{-1}\right\|_\infty \leq 1 - \gamma, \gamma \in (0, 1] \tag{13}$$

$$\Lambda_{\min}(\Sigma_{SS}) \geq C_{\min} > 0 \tag{14}$$

$$\Lambda_{\max}(\Sigma_{SS}) \leq C_{\min} < +\infty \tag{15}$$

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

$$\widehat{\beta} \triangleq \frac{1}{2n} \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \tag{16}$$

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 $\beta_{\min} \triangleq \min_{i \in S} |\beta_i^*| > g(\lambda_n)$, where

$$g(\lambda_n) \triangleq C_5 \lambda_n \left\|\Sigma_{SS}^{-1/2}\right\|_\infty^2 + 20\sqrt{\frac{\sigma^2 \log k}{C_{\min} n}}, \tag{17}$$

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 [2]. 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 [2] 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 [2]. We furnish a sharper substitute in Lemma 2. □

Lemma 2. *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 [2, Lemma 9] (which follows in a straightforward manner from the seminal results of [3]), we know that

$$\mathbb{P}\left[\|A\|_2 \geq \delta(n, k, t)\right] \leq 2e^{-nt^2/2}, \tag{18}$$

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

$$\mathbb{P}\left[\|A\|_2 \geq \frac{c_1}{\sqrt{k}}\right] \leq 2\exp\left\{-\frac{n}{2}\left(\sqrt{\frac{c_1}{\sqrt{k}}} - 1 - \sqrt{\frac{k}{n}}\right)^2\right\} \tag{19}$$

$$\leq 2\exp\left\{-\frac{n}{2}\left(\frac{\sqrt{8}-1}{2}\right)^2\right\} \tag{20}$$

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

$$\delta(n,k,t) = 2\left(\sqrt{\frac{k}{n}} + t\right) + \left(\sqrt{\frac{k}{n}} + t\right)^2 \tag{21}$$

$$\leq \frac{c_1}{\sqrt{k}}. \tag{22}$$

This implies that

$$\mathbb{P}\left[\|A\|_2 \geq \frac{c_1}{\sqrt{k}}\right] \leq 2\exp\left\{-\frac{nc_1^2}{128k}\right\}. \tag{23}$$

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 [2].

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

References

[1] 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.

[2] 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.

[3] 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.