Supplementary Material

"Lower Bounds on Active Learning for Graphical Model Selection" (Scarlett and Cevher, AISTATS 2017)

A Proof of Lemma 1

We start with the following form of Fano's inequality [22, Lemma 1]:

$$1 \geq \frac{\log|\mathcal{T}|}{I(G;\mathbf{X})}\left(1-\delta-\frac{\log 2}{\log|\mathcal{T}|}\right), \tag{43}$$

where $\mathbf{X} = (X^{(1)}, \dots, X^{N})$. This remains valid in the active learning setting since it only relies on the fact that $G \to \mathbf{X} \to \hat{G}$ forms a Markov chain. Despite this common starting point, we bound the mutual information significantly differently. Defining $\mathbf{X}^{(1,i)} = (X^{(1)}, \dots, X^{(i)})$, we have[2]

$$\begin{aligned}
I(G;\mathbf{X}) &= \sum_{i=1}^{N} I(X^{(i)}; G \,|\, \mathbf{X}^{(1,i-1)}) && (44)\\
&= \sum_{i=1}^{N} I(X^{(i)}; G \,|\, \mathbf{X}^{(1,i-1)}, Z^{(i)}) && (45)\\
&= \sum_{i=1}^{N} \Big(H(X^{(i)} \,|\, \mathbf{X}^{(1,i-1)}, Z^{(i)}) - H(X^{(i)} \,|\, \mathbf{X}^{(1,i-1)}, Z^{(i)}, G) \Big) && (46)\\
&= \sum_{i=1}^{N} \Big(H(X^{(i)} \,|\, \mathbf{X}^{(1,i-1)}, Z^{(i)}) - H(X^{(i)} \,|\, Z^{(i)}, G) \Big) && (47)\\
&\leq \sum_{i=1}^{N} \Big(H(X^{(i)} \,|\, Z^{(i)}) - H(X^{(i)} | G, Z^{(i)}) \Big) && (48)\\
&= \sum_{i=1}^{N} I(G; X^{(i)} | Z^{(i)}), && (49)
\end{aligned}$$

where (44) follows from the chain rule, (45) follows since $Z^{(i)}$ is a function of $\mathbf{X}^{(1,i-1)}$, (47) follows since $X^{(i)}$ is conditionally independent of $\mathbf{X}^{(1,i-1)}$ given $(G, Z^{(i)})$, and (48) follows since conditioning reduces entropy. This completes the proof of (11).

Conditioned on $Z^{(i)} = z^{(i)}$, the only variables in $X^{(i)}$ conveying information about G are those corresponding to entries where $z^{(i)}$ is one, since the others deterministically equal $*$. By applying the mutual information upper bound of [22] (see the proof of Corollary 2 therein) to the restricted graph $G(z^{(i)})$ with an auxiliary distribution $Q_{(z^{(i)})}$, we obtain that

$$D(P_{G(z^{(i)})} \| Q_{(z^{(i)})}) \leq \epsilon(z^{(i)}), \forall G \in \mathcal{T} \implies I(G; X^{(i)} | Z^{(i)} = z^{(i)}) \leq \epsilon(z^{(i)}). \tag{50}$$

Note that conditioned on $Z^{(i)} = z^{(i)}$, the graph G may no longer be uniform on $\mathcal{T}$; the preceding claim remains valid since the proof of [22, Cor. 2] is for general graph distributions that need not be uniform.

Finally, the inequality in (12) follows by averaging both sides of the mutual information bound in (50) over $Z^{(i)}$.

B Ensemble and Sample Complexity for Comparing the Average Degree and Maximal Degree (Ising model)

Formalizing the discussion on the Ising model in Section 5, we introduce the following analog of Ensemble 2, consisting of some number L of variable-size cliques with an edge removed.

[2] Here H represents entropy in the discrete case (e.g., Ising), and differential entropy in the continuous case (e.g., Gaussian).

Ensemble2a($m_1, \dots, m_L$) **[Variable-size edge-removed cliques ensemble]:**

- Form L arbitrary node-disjoint cliques of sizes $(m_1, \dots, m_L)$, to obtain a base graph G'.
- Each graph in $\mathcal{T}$ is obtained by removing a single edge from each of the L cliques.

We have the following.

Lemma 2. *Fix the integers L and $(m_1, \dots, m_L)$ with $\sum_{j=1}^{L} m_j = p$, and let G be drawn uniformly from* Ensemble2a*$(m_1, \dots, m_L)$. Then in order to achieve $\overline{P}_{\rm e} \le \delta$, it is necessary that*

$$n \geq \frac{e^{\lambda d_{\max}} \log\big(d_{\max}(d_{\max}+1)\big)}{2\lambda d_{\max} e^{\lambda}} \Big(1 - \delta - \frac{\log 2}{\log(d_{\max}+1)}\Big), \tag{51}$$

where $d_{\max} = \max_{j=1,\dots,L} m_j - 1$.

Proof. We consider a genie argument, in which the decoder is informed of all of the removed edges from the cliques, except for the largest, whose size is $d_{\max}+1$. In this case, the analysis reduces to that of Ensemble2($d_{\max}+1$) on a graph with $p = d_{\max}+1$ nodes. The result now follows immediately from (20), and recalling that the $o(1)$ remainder term therein is equal to $\frac{\log 2}{|\mathcal{T}|}$ from (11). □

C Ensemble and Sample Complexity for Comparing the Average Degree and Maximal Degree (Gaussian model)

Formalizing the discussion on the Gaussian model in Section 5, we introduce the following ensemble, consisting of some number L of variable-size cliques.

Ensemble4a($m_1, \dots, m_L$) **[Disjoint variable-size cliques ensemble]:**

- Each graph in $\mathcal{T}$ consists of L disjoint cliques of sizes $(m_1, \dots, m_L)$ nodes that may otherwise be arbitrary.

We have the following.

Lemma 3. *Fix the integers L and $(m_1, \dots, m_L)$ with $\sum_{j=1}^{L} m_j = p$ and $\max_{j=1,\dots,L} m_j = o(p)$, and let G be drawn uniformly from* Ensemble4a*$(m_1, \dots, m_L)$. Then for any $\alpha \in (0,1)$ (not depending on p), in order to achieve $\overline{P}_{\rm e} \le \delta$, it is necessary that*

$$n \geq \frac{2\alpha p d_{\min}^{(\alpha)} \log \frac{p}{d_{\max}}}{\log\Big(1 + \big((d_{\max}+1)\frac{\tau}{1-\tau}\big)^2\Big)} \big(1 - \delta - o(1)\big), \tag{52}$$

where $d_{\max} = \max_{j=1,\dots,L} m_j - 1$, and $d_{\min}^{(\alpha)}$ is the minimum degree among the αp nodes having the largest degree.[3]

Proof. We again consider a genie argument, in which the decoder is informed of all of the cliques except the largest ones, such that these remaining cliques form a total of αp nodes.[4] Assuming without loss of generality that the m_j are in decreasing order, the analysis reduces to the study of Ensemble4a on a graph with αp nodes, and cliques of size $(m_1, \dots, m_{L'})$, where $L' \le L$ is defined such that $\sum_{j=1}^{L'} m_j = \alpha p$.

[3] This is the same for all graphs in the ensemble, so here $d_{\min}^{(\alpha)}$ is well-defined.

[4] Since $m_j = o(p)$ for all j, we can safely ignore rounding and assume that the total is exactly αp.

For this reduced ensemble, the total number of graphs is $\binom{\alpha p}{m_1}\binom{\alpha p - m_1}{m_2}\dots\binom{\alpha p - \sum_{j=1}^{L'-2} m_j}{m_{L'-1}}\binom{m_{L'}}{m_{L'}}$. We let L'' be the largest integer such that $\sum_{j=1}^{L''} m_j \le \alpha p/2$, and write

$$\log|\mathcal{T}| \ge \sum_{j=1}^{L''} \log\binom{\lfloor \alpha p/2 \rfloor}{m_j} \tag{53}$$

$$= \sum_{j=1}^{L''} \Big(m_j \log \frac{\alpha p}{2m_j} \Big)(1+o(1)) \tag{54}$$

$$\ge \Big(\frac{\alpha p}{2} \log \frac{\alpha p}{2m_1} \Big)(1+o(1)) \tag{55}$$

$$= \Big(\frac{\alpha p}{2} \log \frac{p}{d_{\max}} \Big)(1+o(1)), \tag{56}$$

where (54) follows since $m_j = o(\alpha p)$ by assumption, (55) follows by first applying $m_j \le m_1$ inside the logarithm and then applying the definition of L'', and (56) follows since $m_1 = d_{\max}+1$ by definition.

We now follow the analysis of Section 4.3.2, and note that if a single measurement consists of $n(z)$ nodes indexed by $z \in \{0,1\}^p$, and if this corresponds to observing $\widetilde{m}_j$ nodes from each clique $j = 1,\dots,L'$, then we have the following analog of (34):

$$D(P_{G(z)} \| Q_{(z)}) = \frac{1}{2}\Bigg(\sum_{j=1}^{L'} -\log\Big(1 - \frac{\widetilde{m}_j a}{1+m_j a}\Big) - \frac{\widetilde{m}_j a}{1+m_j a} \Bigg), \tag{57}$$

where $Q_{(z)}$ and a are defined in Section 4.3.2.

Defining $\beta_j = \frac{\widetilde{m}_j a}{1+m_j a}$ and $f(\beta) = \frac{-\log(1-\beta)-\beta}{\beta}$, we can write the right-hand side of (57) as

$$\sum_{j=1}^{L'} \beta_j f(\beta_j), \tag{58}$$

As a result, we consider the maximization of (35) subject to $0 \le \beta_j \le \frac{m_j a}{1+m_j a}$ and $\sum_j \beta_j(1+m_j a) = n(z)a$, where these constraints follow immediately from $0 \le \widetilde{m}_j \le m$ and $\sum_j \widetilde{m}_j = n(z)$.

While the optimal choices of $\{\beta_j\}$ for the preceding maximization problem are unclear, we observe that the final objective value can only increase if we relax the second constraint to $\sum_j \beta_j(1+m_{\min}^{(\alpha)} a) \le n(z)a$, where $m_{\min}^{(\alpha)} = m_{L'} = d_{\min}^{(\alpha)} + 1$. With this modification, we find similarly to (34) that the maximum is achieved by setting β_j to its maximum value $\frac{m_j a}{1+m_j a}$ (i.e., $\widetilde{m}_j = m_j$) for as many of the largest cliques as is permitted by the constraint $\sum_j \beta_j(1+m_{\min}^{(\alpha)} a) \le n(z)a$. Since each clique under consideration has at least $m_{\min}^{(\alpha)}$ nodes, this amounts to at most $\frac{n(z)}{m_{\min}^{(\alpha)}}$ cliques. Moreover, since $\beta f(\beta)$ is increasing in β, the corresponding values of $\beta_j f(\beta_j)$ are upper bounded by $\frac{m_{\max} a}{1+m_{\max} a} f\big(\frac{m_{\max} a}{1+m_{\max} a}\big)$.

Combining these observations, we obtain the following analog of (37):

$$\sum_{j=1}^{L'} \beta_j f(\beta_j) \le \frac{n(z)}{m_{\min}^{(\alpha)}} \frac{m_{\max} a}{1+m_{\max} a} f\Big(\frac{m_{\max} a}{1+m_{\max} a}\Big), \tag{59}$$

and accordingly, using the same steps to those following (34), we obtain the following analog of (41):

$$\sum_{i=1}^{N} I(G; X^{(i)} | Z^{(i)}) \le \frac{n}{4 m_{\min}^{(\alpha)}} \log\big(1 + (m_{\max} a)^2\big). \tag{60}$$

The proof is concluded using (11) along with the cardinality bound in (56), and recalling that $m_{\min}^{(\alpha)} = d_{\min}^{(\alpha)} + 1$, $m_{\max} = d_{\max}+1$, and $a = \frac{\tau}{1-\tau}$.

□