Lower Bounds on Active Learning for Graphical Model Selection

Jonathan Scarlett and Volkan Cevher
Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
Email: {jonathan.scarlett, volkan.cevher}@epfl.ch

Abstract

We consider the problem of estimating the underlying graph associated with a Markov random field, with the added twist that the decoding algorithm can iteratively choose which subsets of nodes to sample based on the previous samples, resulting in an active learning setting. Considering both Ising and Gaussian models, we provide algorithm-independent lower bounds for high-probability recovery within the class of degree-bounded graphs. Our main results are minimax lower bounds for the active setting that match the best known lower bounds for the passive setting, which in turn are known to be tight in several cases of interest. Our analysis is based on Fano's inequality, along with novel mutual information bounds for the active learning setting, and the application of restricted graph ensembles. While we consider ensembles that are similar or identical to those used in the passive setting, we require different analysis techniques, with a key challenge being bounding a mutual information quantity associated with observed subsets of nodes, as opposed to full observations.

1 Introduction

Graphical models are a widely-used tool for providing compact representations of the conditional independence relations between random variables, and arise in areas such as image processing [1], statistical physics [2], computational biology [3], natural language processing [4], and social network analysis [5]. The problem of *graphical model selection* consists of recovering the graph structure given a number of independent samples from the underlying distribution. While this problem is NP-hard in general [6], there exist a variety of methods guaranteeing exact recovery with high probability on *restricted* graph classes, with a particularly common restriction being bounded degree.

Several variations of graphical model selection problems with *active learning* have appeared in the literature. In this paper, we adopt the formulation given in [7], in which the recovery algorithm may adaptively choose which nodes to sample, based on the previous samples. The goal is to recover the underlying graph subject to a constraint on the total number of node observations. As discussed in [7], this variation is of interest in several applications; for example, in sensor networks one may be able to choose which sensors to activate, rather than simultaneously activating every sensor at every time instant. Only upper bounds were provided in [7], and the problem of finding lower bounds was left as an open problem.

1.1 Contributions

In this paper, we complement the work of [7] by providing algorithm-independent lower bounds on active learning for graphical model selection. Our main findings are summarized as follows:

1. For both Ising models and Gaussian models, we provide lower bounds that essentially match the best known lower bounds for the passive setting [8, 9], in terms of the minimax probability of error with respect to the class of bounded-degree graphs. The passive learning bounds are known to be tight in several cases of interest, and our results show that *active learning does not help significantly in the minimax sense* in such cases.
2. We provide a class of Gaussian graphical models where the average degree dictates the lower bounds as opposed to the maximal degree, and where we match upper bounds based on the average degree in [7]. Hence, we identify a graph class

Appearing in Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Flordia, USA. JMLR: W&CP volume 54.

where the average degree is provably the fundamental quantity dictating the fundamental limits. Moreover, we provide a class of Ising models where the maximal degree provably remains the key quantity dictating the performance, hence revealing that one cannot always improve the dependence from the maximal to the average degree.

Our analysis uses a variation of Fano's inequality for the active learning setting, along with novel mutual information bounds proved using analogous techniques to those used in channel coding with noiseless feedback [10]. We apply the resulting bound to a variety of restricted graph ensembles in which the graphs are difficult to distinguish from each other, with notable examples being (i) isolated edges that are difficult to detect; (ii) cliques with a single edge removed such that the removal is difficult to detect. While the ensembles that we use are similar or identical to those used in the passive setting, analyzing them in the active setting requires new techniques, particularly for bounding a mutual information quantity associated with partial observations instead of full observations.

1.2 Related Work

In the same way that feedback often provides little or no gain in the capacity for channel coding [10], it is often observed that active learning provides little or no gain in the information-theoretic sample complexity of inference and learning problems. For example, in the compressive sensing problem, it has been shown that the improvement amounts to at most a logarithmic factor [11]. For the group testing problem, under a broad range of scalings of the sparsity level, not even the constant factors improve [12, 13].

On the other hand, active learning is known to strictly improve the sample complexity in several cases of interest [14, 15]. Moreover, it should be noted that even when adaptivity does not help asymptotically in an information-theoretic sense, it can still help in the sense of leading to simpler and less computationally expensive algorithms, and also in improving the non-asymptotic performance [13, 15, 16].

Active learning for graphical model selection has been studied in several contexts [7, 17, 18], the most relevant to ours being that of Dasarathy *et al.* [7]. A general algorithm was proposed therein using abstract subroutines for neighborhood selection and neighborhood verification, and applications to the Gaussian setting revealed cases where the total number of node observations is improved from $O(d_{\max} p \log p)$ to $O((1+\overline{d}_{\max})p \log p)$. Here $\overline{d}_{\max}$ is the average of the node-wise maximal degree, where the latter is defined as the highest degree among a node and all its neighbors. This quantity can be significantly smaller than $d_{\max}$, in which case the improvement in the sample complexity is substantial.

Information-theoretic lower bounds for the passive setting were given in [8, 19–24] for the Ising model, and [9, 25, 26] for the Gaussian model. Let $\widetilde{n}$ be the sample complexity with respect to the number of p-dimensional observations. The best minimax lower bounds for degree-bounded graphs are summarized as follows for the Ising model [8]:

$$\widetilde{n} = \Omega\bigg(\max\bigg\{ \frac{\log p}{\lambda \tanh \lambda}, \frac{e^{\lambda d}\log(pd)}{\lambda d e^{\lambda}}, d \log \frac{p}{d} \bigg\} \bigg), \quad (1)$$

where p is the number of nodes, d is the maximal degree, and λ is the inverse temperature of the Ising model (see Section 2 for precise definitions). For the Gaussian model, the best known lower bounds for degree-bounded graphs are [9]

$$\widetilde{n} = \Omega\bigg(\max\bigg\{ \frac{\log p}{\tau^2}, \frac{d \log \frac{p}{d}}{\log(1+d\tau)} \bigg\} \bigg), \quad (2)$$

where τ corresponds to the smallest allowed off-diagonal magnitude in the normalized inverse covariance matrix (see Section 2 for details).

A wide range of polynomial-time algorithms have been proposed for the passive learning of graphical models; see [19, 20, 24, 27–32] for Ising models, and [20, 33–35] for Gaussian models. The best performance bounds among these algorithms match those of (1)–(2) in several cases of interest, though there are other cases where gaps remain, or where the results are difficult to compare due to the differences in the underlying assumptions (e.g., additional coherence assumptions).

1.3 Structure of the Paper

In Section 2, we formally define the Ising and Gaussian graphical models, and formulate the active learning problem. Our main results are presented and discussed in Section 3. The proofs are given in Section 4.1 (Fano's inequality), Section 4.2 (Ising model), and Section 4.3 (Gaussian model). In Section 5, we discuss the role of the average vs. maximal degree, and we conclude our work in Section 6.

2 Active Learning for Graphical Model Selection

2.1 Preliminaries

We consider a collection of p random variables $(X_1, \ldots, X_p)$ whose joint distribution is encoded by a graphical model $G = (V, E)$ with vertex set $V =$

$\{1,\dots,p\}$ and undirected edge set E. The elements of V are referred to as *nodes* or *variables* interchangeably. We use the standard terminology that the *degree* of a node $i \in V$ is the number of edges in E containing i, and that a *clique* is a fully-connected subset of V of cardinality at least two.

We consider two classes of joint probability distributions encoded by G, namely, Ising models and Gaussian models. These are described as follows.

Ising Model: In the ferromagnetic Ising model [36,37], each vertex is associated with a binary random variable $X_i \in \{-1,1\}$, and the corresponding joint distribution is described by the probability mass function

$$P_G(x) = \frac{1}{Z}\exp\left(\lambda \sum_{(i,j)\in E} x_i x_j\right), \tag{3}$$

where Z is a normalizing constant called the partition function. Here $\lambda > 0$ is a parameter to the distribution, sometimes called the inverse temperature.

In the context of Ising model selection, we write $\mathcal{G}_d$ as $\mathcal{G}_{d,\lambda}$ to emphasize that the results depend on λ. Although we let λ be a constant here, our lower bounds remain valid in the minimax sense when one considers the larger class in which the edges have differing parameters $\{\lambda_{ij}\}$ in the range $[\lambda_{\min},\lambda_{\max}]$, provided that $\lambda_{\min} \le \lambda \le \lambda_{\max}$.

Gaussian Model: In the Gaussian graphical model [37], each vertex is associated with a random variable $X_i \in \mathbb{R}$, and the corresponding joint distribution is

$$(X_1,\dots,X_p) \sim N(\mathbf{0},\mathbf{\Sigma}), \tag{4}$$

where $\mathbf{0}$ is the vector of zeros, and $\mathbf{\Sigma}$ is a covariance matrix whose inverse $\mathbf{\Sigma}^{-1}$ contains non-zeros only in the diagonal entries and the indices corresponding to pairs in E. By the Hammersley-Clifford theorem [37], this implies the Markov property for the graph, namely, that a given node is conditionally independent of the rest of the graph given its neighbors. The joint density function corresponding to (4) is denoted by P_G, overloading the notation used above for the Ising model.

A typical restriction on the entries of $\Theta = \Sigma^{-1}$ is that $\frac{|\Theta_{ij}|}{\sqrt{\Theta_{ii}\Theta_{jj}}}$ is lower bounded by some constant $\tau > 0$ [7,9]. We consider the simplest special case of this in which the lower bound always holds with equality:

$$\Theta_{ij} = \begin{cases} 1 & i = j \\ \pm\tau & (i,j) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

We write $\mathcal{G}_d$ as $\mathcal{G}_{d,\tau}$ to emphasize that the results depend on τ. Similarly to the Ising model, our lower bounds remain valid in the minimax case when we consider the larger class with $\frac{|\Theta_{ij}|}{\sqrt{\Theta_{ii}\Theta_{jj}}} \in [\tau_{\min},\tau_{\max}]$ with $\tau_{\min} \le \tau \le \tau_{\max}$.

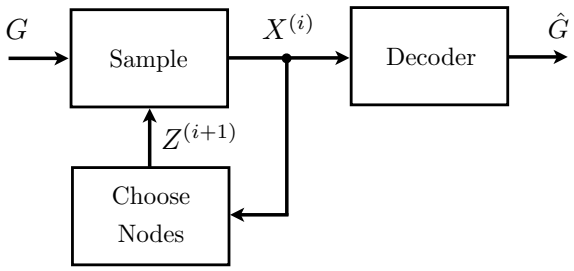

Figure 1: Illustration of the active learning problem for graphical model selection.

2.2 Problem Statement

The problem of graphical model selection with active learning proceeds in rounds $i = 1,2,\dots,$ as illustrated in Figure 1. In the i-th round, the algorithm selects a subset of V to observe, encoded by a binary vector $Z^{(i)} \in \{0,1\}^p$ equaling one for observed nodes and zero for non-observed nodes. The resulting *sample* (or *observation*) is a p-dimensional vector $X^{(i)}$ such that:

- The joint distribution of the entries of $X^{(i)}$, corresponding to the entries where $Z^{(i)}$ is one, coincide with the corresponding joint distribution of the vector $(X_1,\dots,X_p) \sim P_G$, with independence between rounds;
- The values of the entries of $X^{(i)}$, corresponding to the entries where $Z^{(i)}$ is zero, are deterministically given by $*$, a symbol indicating that the node was not observed.

For convenience, we let N denote the maximum possible number of active learning rounds (e.g., we can simply set $N = n$), and use the convention that for values of i beyond the actual (possibly random) final round, $X^{(i)} = (*,\dots,*)$. Letting $|Z^{(i)}|$ denote the number of entries where $Z^{(i)}$ is one, we refer to $\sum_{i=1}^{N}|Z^{(i)}|$ as the *total number of node observations* used throughout the course of the algorithm, and we impose an upper bound on its maximum allowed value, denoted by n. Note that this differs from the quantity $\widetilde{n}$ in (1)–(2) by a factor of p.

After the final round, the algorithm constructs an estimate $\hat{G}$ of G, and the error probability is given by

$$P_{\mathrm{e}}(G) := \mathbb{P}[\hat{G} \neq G]. \tag{6}$$

We consider the class $\mathcal{G}_d$ of degree-bounded graphs, in which all nodes have degree at most d. Specifically, we

are interested in bounds on the minimax (worst-case) error probability for graphs in this class:

$$P_{\mathrm{e}} := \max_{G \in \mathcal{G}_d} \mathbb{P}[\hat{G} \neq G], \tag{7}$$

where the dependence on the total number of node samples n is kept implicit. Note that when we consider the Gaussian setting, the maximum in (6) is not only over the graph G, but also implicitly over the signs ($+1$ or -1) in the second case of (5).

We are interested in characterizing the *sample complexity*, meaning the required number of node observations n needed in order to achieve $P_{\mathrm{e}} \leq \delta$ for some target error probability $\delta > 0$.

3 Main Results

In this section, we state and discuss our main results, namely, minimax lower bounds on the sample complexity for $\mathcal{G}_d$. We note that the proofs are based on graph ensembles in which the maximal degree and average degree are approximately equal; however, in Section 5, we discuss variations of these ensembles in which these two notions differ significantly.

3.1 Ising Model

Theorem 1. *For Ising graphical models with $\lambda d \geq 1$, in order to recover any graph in $\mathcal{G}_{d,\lambda}$ with probability at least $1-\delta$, it is necessary that the total number of node observations, n, satisfies*

$$n \geq \max\left\{ \frac{2p\log p}{\lambda \tanh \lambda}, \frac{e^{\lambda d}\log(pd)}{2\lambda d e^{\lambda}}, \frac{pd \log \frac{p}{8d}}{4\log 2} \right\} \times (1-\delta-o(1)). \tag{8}$$

Proof. See Section 4.2. □

The second bound in (8) reveals that the sample complexity is very large when $\lambda d \to \infty$ at a rate that is not too slow, due to the exponential term $e^{\lambda d}$. On the other hand, when $\lambda = O(\frac{1}{d})$, the first bound gives a sample complexity of $\Omega(d^2 p \log p)$, since $\tanh \lambda = O(\lambda)$ as $\lambda \to 0$. Finally, in any case, the third bound gives $n = \Omega\big(pd \log \frac{p}{d}\big)$. These observations coincide with those for the lower bounds on passive learning in [8] (see (1) with $\widetilde{n} = np$), suggesting that active learning does not help much in the minimax sense for $\mathcal{G}_{d,\lambda}$. Note that compared to [8], we lose a factor of p in the second bound, but this factor is insignificant compared to $e^{\lambda d}$ provided that $\lambda d \gg \log p$.

3.2 Gaussian Model

Theorem 2. *For Gaussian graphical models with $d = o(p)$, in order to recover any graph in $\mathcal{G}_{d,\tau}$ with probability at least $1-\delta$, it is necessary that the total number of node observations, n, satisfies*

$$n \geq \max\left\{ \frac{4p\log p}{\log \frac{1}{1-\tau^2}}, \frac{2pd\log \frac{p}{d}}{\log\Big(1+\big((d+1)\frac{\tau}{1-\tau}\big)^2\Big)} \right\} \times (1-\delta-o(1)). \tag{9}$$

Proof. See Section 4.3. □

When $\tau = o(1)$, the first bound behaves as $\Omega\big(\frac{1}{\tau^2} p \log p\big)$, whereas when τ is a constant, the second bound behaves as $\Omega\big(\frac{1}{\log d} \cdot pd \log \frac{p}{d}\big)$. Both of these scaling laws are identical to the necessary conditions for passive learning in [9] (see (2) with $\widetilde{n} = np$), again suggesting that active learning does not help much in the minimax sense for $\mathcal{G}_{d,\tau}$.

While the above findings indicate that active learning does not help much in the minimax sense for $\mathcal{G}_d$, we discuss a more restricted class of graphs in Section 5 for which active learning helps when τ is a constant. Specifically, similarly to the upper bound in [7], the linear dependence on the maximal degree d in the second term of (9) is improved to the *average* degree.

4 Proofs of Main Results

4.1 Fano's Inequality for Active Learning

We first apply Fano's inequality [10] along with a novel mutual information bound for active learning in graphical model selection. The proof bears some resemblance to that of the converse bound for channel coding with noiseless feedback [10, Sec. 7.12].

For $z \in \{0,1\}^p$, we let $G(z)$ denote the subgraph of G obtained by keeping only the nodes corresponding to entries where z equals one, and denote the resulting joint distribution by $P_{G(z)}$. More generally, for a joint distribution Q on p random variables labeled $\{1, \cdots, p\}$, we let $Q_{(z)}$ denote the joint marginal distribution corresponding to the entries where z is one.

In the following lemma, we let G be uniformly random on some subset of $\mathcal{G}_d$, and define the *average* error probability

$$\overline{P}_{\mathrm{e}} := \mathbb{P}[\hat{G} \neq G] = \mathbb{E}[P_{\mathrm{e}}(G)], \tag{10}$$

where in contrast with (7), the probability is now additionally over G. Clearly any lower bound on the sample complexity for achieving $\overline{P}_{\mathrm{e}} \leq \delta$ implies the

same lower bound for achieving $P_{\mathrm{e}} \leq \delta$, since P_{e} is defined with respect to the worst case.

Lemma 1. *Let G be uniform over a restricted graph class $\mathcal{T} \subseteq \mathcal{G}_d$. In order to achieve $\overline{P}_{\mathrm{e}} \leq \delta$, it is necessary that*

$$1 \geq \frac{\log|\mathcal{T}|}{\sum_{i=1}^{N} I(G; X^{(i)}|Z^{(i)})}\left(1 - \delta - \frac{\log 2}{\log|\mathcal{T}|}\right), \quad (11)$$

where N is the maximum possible number of active learning rounds. Moreover, if there exists a p-dimensional joint distribution Q such that $D(P_{G(z)}\|Q_{(z)}) \leq \epsilon(z)$ for all $G \in \mathcal{T}$ and $z \in \{0,1\}^p$, where $\epsilon(z)$ is some non-negative function, then we have

$$I(G; X^{(i)}|Z^{(i)}) \leq \mathbb{E}\big[\epsilon(Z^{(i)})\big] \quad (12)$$

for all i.

The proof is given in the supplementary material. The high-level steps are as follows: (i) Bound the error probability in terms of $I(G;\mathbf{X})$ using Fano's inequality; (ii) Use the chain rule to write $I(G;\mathbf{X}) = \sum_{i=1}^{N} I(X^{(i)}; G \,|\, X^{(1)}, \dots, X^{(i-1)})$; (iii) Upper bound the summands via analogous steps to the proof of the channel coding theorem with feedback [10, Sec. 7.12]; (iv) Relate the divergence $D(P_{G(z)}\|Q_{(z)})$ to $I(G; X^{(i)}|Z^{(i)})$ using similar steps to [22].

4.2 Proof of Theorem 1 (Ising model)

4.2.1 First Bound for the Ising Model

We use the following ensemble in which every node has degree one.

Ensemble1 [Isolated edges ensemble]

- Each graph in $\mathcal{T}$ consists of $\lfloor p/2 \rfloor$ node-disjoint edges that may otherwise be arbitrary.

The total number of graphs is $|\mathcal{T}| = \binom{p}{2}\binom{p-2}{2}\dots\binom{4}{2}\binom{2}{2}$ (or similarly when p is an odd number), which is lower bounded by $\binom{\lfloor p/2 \rfloor}{2}^{\lfloor p/2 \rfloor}$, yielding

$$\log|\mathcal{T}| \geq \left\lfloor \frac{p}{2} \right\rfloor \log \binom{\lfloor p/2 \rfloor}{2} = (p \log p)(1 + o(1)). \quad (13)$$

To obtain a mutual information bound of the form (12), we choose $Q = P_{G'}$ with G' being the empty graph, and note that for a fixed $z \in \{0,1\}^p$ containing $n(z)$ ones, $G(z)$ consists of at most $n(z)/2$ node-disjoint edges. Since the divergence corresponding to graphs differing in a single edge is upper bounded by $\lambda \tanh \lambda$ [22], and since the divergence is additive for independent products, we obtain $D(P_{G(z)}\|P_{G'(z)}) \leq \frac{n(z)}{2}\lambda \tanh\lambda$, and hence (12) becomes

$$I(G; X^{(i)}|Z^{(i)}) \leq \frac{1}{2}\mathbb{E}[n(Z^{(i)})]\lambda \tanh\lambda. \quad (14)$$

Summing over i and noting that $\sum_{i=1}^{N} n(Z^{(i)}) \leq n$ with probability one, since the algorithm can only use up to n node observations, we obtain

$$\sum_{i=1}^{N} I(G; X^{(i)}|Z^{(i)}) \leq \frac{n}{2}\lambda \tanh\lambda. \quad (15)$$

Substitution into (11) yields the necessary condition

$$n \geq \frac{2p \log p}{\lambda \tanh\lambda}(1 - \delta - o(1)), \quad (16)$$

where the numerator arises from (13).

4.2.2 Second Bound for the Ising Model

We use the following ensemble from [8].

Ensemble2(m) [Clique-minus-one ensemble]:

- Form $\lfloor \frac{p}{m} \rfloor$ arbitrary node-disjoint cliques containing m nodes each, to form a base graph G'.
- Each graph in $\mathcal{T}$ is obtained by removing a single edge from G'.

We choose $m = d+1$, so that the maximal degree is d. The total number of graphs is $\lfloor \frac{p}{m} \rfloor \binom{m}{2}$, which yields

$$\log|\mathcal{T}| = (\log(pd))(1 + o(1)). \quad (17)$$

We obtain a bound of the form (12) by choosing $Q = P_{G'}$ with G' as in the ensemble definition. The divergence associated with the full graphs satisfies $D(P_G\|P_{G'}) \leq \frac{4\lambda d e^{\lambda}}{e^{\lambda d}}$ when $\lambda d \geq 1$ [8, Lemma 2]. Since $G(z)$ and $G'(z)$ are common subgraphs of G and G', we trivially have $D(P_{G(z)}\|P_{G'(z)}) \leq D(P_G\|P_{G'})$, and hence $D(P_{G(z)}\|P_{G'(z)})$ satisfies the same upper bound as $D(P_G\|P_{G'})$ regardless of z. Hence, (12) yields

$$I(G; X^{(i)}|Z^{(i)}) \leq \frac{4\lambda d e^{\lambda}}{e^{\lambda d}}. \quad (18)$$

Since the node observation budget is n, the active learning can be done in at most $n/2$ rounds without loss of optimality (i.e., excluding trivial cases where only one node is observed), and we have

$$\sum_{i=1}^{N} I(G; X^{(i)}|Z^{(i)}) \leq \frac{2n\lambda d e^{\lambda}}{e^{\lambda d}}. \quad (19)$$

Substitution into (11) yields the necessary condition

$$n \geq \frac{e^{\lambda d}\log(pd)}{2\lambda d e^{\lambda}}(1 - \delta - o(1)), \quad (20)$$

where the numerator arises from (17).

4.2.3 Third Bound for the Ising Model

We use the following straightforward ensemble, which was also used in [8].

Ensemble3 [Complete ensemble]:

- $\mathcal{T}$ contains all graphs with maximal degree at most d, i.e., $\mathcal{T} = \mathcal{G}_d$.

It was shown in [8] that $\log|\mathcal{T}| \geq \frac{dp}{4}\log\frac{p}{8d}$. To bound the mutual information in (11), we note that the following holds when $z^{(i)}$ contains $n(z^{(i)})$ ones, and hence $n(z^{(i)})$ nodes are observed in the i-th round:

$$I(G; X^{(i)}|Z^{(i)} = z^{(i)}) \leq n(z^{(i)})\log 2. \tag{21}$$

This is because the remaining $p - n(z^{(i)})$ nodes are deterministically equal to $*$, whereas the $n(z^{(i)})$ nodes are binary and hence reveal at most $\log 2$ bits of information each. Summing (21) over i and averaging over $Z^{(i)}$, we obtain

$$\sum_{i=1}^{N} I(G; X^{(i)}|Z^{(i)}) \leq n\log 2, \tag{22}$$

and substitution into (11) yields the desired result.

4.3 Proof of Theorem 2 (Gaussian model)

4.3.1 First Bound for the Gaussian Model

We re-use Ensemble 1 above and apply the same analysis, with the only difference being the bounding of the divergence $D(P_{G_1}\|P_{G_0})$ when G_1 contains one edge and G_0 contains no edges.

When an edge is present, we let the resulting 2×2 covariance matrix and its inverse be given by

$$\mathbf{\Sigma}_1 = (1-\tau^2)\begin{bmatrix} 1 & \tau \\ \tau & 1 \end{bmatrix}, \quad \mathbf{\Sigma}_1^{-1} = \begin{bmatrix} 1 & -\tau \\ -\tau & 1 \end{bmatrix}, \tag{23}$$

whereas for the graph without the edge we simply have $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_0^{-1} = \mathbf{I}$. Both of these choices are clearly consistent with (5).

The divergence between two zero-mean Gaussian vectors of dimension k is

$$D(P_1\|P_0) = \frac{1}{2}\bigg(\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_1) - k + \log\frac{\det\mathbf{\Sigma}_0}{\det\mathbf{\Sigma}_1}\bigg), \tag{24}$$

and with the above covariance matrices and $k = 2$, this simplifies to

$$D(P_1\|P_0) = \frac{1}{2}\log\frac{1}{1-\tau^2}. \tag{25}$$

Hence, in analogy with (16), we obtain

$$n \geq \frac{4p\log p}{\log\frac{1}{1-\tau^2}}(1-\delta-o(1)). \tag{26}$$

4.3.2 Second Bound for the Gaussian Model

We make use of the following ensemble that is similar to one in [9], but with multiple cliques as opposed to only a single one. It can also be thought of as a generalization of Ensemble 1, which corresponds to $m = 2$.

Ensemble4(m) [Disjoint cliques ensemble]:

- Each graph in $\mathcal{T}$ consists of $\lfloor\frac{p}{m}\rfloor$ disjoint cliques of m nodes that may otherwise be arbitrary.

The total number of graphs is $\binom{p}{m}\binom{p-m}{m}\ldots\binom{2m}{m}\binom{m}{m}$ (or analogously when p does not divide m), which is lower bounded by $\binom{\lfloor p/2\rfloor}{m}^{\frac{1}{2}\lfloor\frac{p}{m}\rfloor}$, yielding

$$\log|\mathcal{T}| \geq \frac{1}{2}\Big\lfloor\frac{p}{m}\Big\rfloor\log\binom{\lfloor p/2\rfloor}{m} = \bigg(\frac{p}{2}\log\frac{p}{m}\bigg)(1+o(1)) \tag{27}$$

assuming that $m = o(p)$ and hence $\log\binom{\lfloor p/2\rfloor}{m} = \big(m\log\frac{p}{m}\big)(1+o(1))$. We choose $m = d+1$ so that the maximal degree is d, yielding

$$\log|\mathcal{T}| \geq \bigg(\frac{p}{2}\log\frac{p}{d}\bigg)(1+o(1)). \tag{28}$$

As in [9], we let the inverse covariance matrix associated with a single clique be given by

$$\mathbf{\Sigma}_1^{-1} = \begin{bmatrix} 1+a & a & \cdots & a \\ a & 1+a & \cdots & a \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \cdots & 1+a \end{bmatrix}, \tag{29}$$

for $a > 0$, yielding a covariance matrix given by

$$\mathbf{\Sigma}_1 = \frac{1}{1+ma} \times \begin{bmatrix} 1+(m-1)a & -a & \cdots & -a \\ -a & 1+(m-1)a & \cdots & -a \\ \vdots & \vdots & \ddots & \vdots \\ -a & -a & \cdots & 1+(m-1)a \end{bmatrix} \tag{30}$$

We set $a = \frac{\tau}{1-\tau}$ to ensure that the ratio of off-diagonals to diagonals in $\mathbf{\Sigma}_1^{-1}$ is τ, in accordance with (5). Note that this form of the inverse covariance matrix is slightly different to that in (5), but the difference only amounts to scaling all observations by a factor of

$\sqrt{1+a}$, and hence the recovery problem is unchanged regardless of which form is assumed.

To obtain a bound of the form (12), we let Q be jointly Gaussian with mean zero and identity covariance matrix, defining $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_0^{-1} = \mathbf{I}$ accordingly. We first study the behavior of the divergence $D(P_{G(z)}\|Q_{(z)})$ when all of the non-zero values of z correspond to nodes within a single clique in G. Hence, z contains $\widetilde{m} \in \{1,\dots,m\}$ non-zero entries.

Letting $\widetilde{\mathbf{\Sigma}}_1$ denote an arbitrary sub-matrix of $\mathbf{\Sigma}_1$ corresponding to $\widetilde{m} \in \{1,\dots,m\}$ nodes, a straightforward computation gives

$$\det \widetilde{\mathbf{\Sigma}}_1 = \frac{1+(m-\widetilde{m})a}{1+ma} = 1 - \frac{\widetilde{m}a}{1+ma} \tag{31}$$

$$\mathrm{Tr}(\widetilde{\mathbf{\Sigma}}_1) = \widetilde{m}\frac{1+(m-1)a}{1+ma} = \widetilde{m}\bigg(1 - \frac{a}{1+ma}\bigg). \tag{32}$$

Defining $\widetilde{\mathbf{\Sigma}}_0$ analogously simply gives $\widetilde{\mathbf{\Sigma}}_0 = \widetilde{\mathbf{\Sigma}}_0^{-1} = \mathbf{I}$, and hence (24) with $k = \widetilde{m}$ gives

$$D(\widetilde{P}_1\|\widetilde{P}_0) = \frac{1}{2}\bigg(-\log\bigg(1-\frac{\widetilde{m}a}{1+ma}\bigg) - \frac{\widetilde{m}a}{1+ma}\bigg) \tag{33}$$

for $\widetilde{P}_0 \sim N(\mathbf{0}, \widetilde{\mathbf{\Sigma}}_0)$ and $\widetilde{P}_1 \sim N(\mathbf{0}, \widetilde{\mathbf{\Sigma}}_1)$.

Suppose now that a single measurement consists of $n(z)$ nodes indexed by $z \in \{0,1\}^p$. For a fixed graph $G \in \mathcal{T}$, this amounts to observing $\widetilde{m}_j$ nodes from each clique $j = 1,\dots,\lfloor \frac{p}{m} \rfloor$, for some integers $\{\widetilde{m}_j\}$ such that $\sum_{j=1}^{\lfloor \frac{p}{m} \rfloor} \widetilde{m}_j = n(z)$. Since the divergence is additive for independent products, we obtain

$$D(P_{G(z)}\|Q_{(z)}) = \frac{1}{2}\Bigg(\sum_{j=1}^{\lfloor \frac{p}{m} \rfloor} -\log\bigg(1-\frac{\widetilde{m}_j a}{1+ma}\bigg) - \frac{\widetilde{m}_j a}{1+ma}\Bigg). \tag{34}$$

To simplify the subsequent exposition, we write the summation as

$$\sum_{j=1}^{\lfloor \frac{p}{m} \rfloor} \beta_j f(\beta_j), \tag{35}$$

where $\beta_j = \frac{\widetilde{m}_j a}{1+ma}$ and $f(\beta) = \frac{-\log(1-\beta)-\beta}{\beta}$. We consider the maximization of (35) subject to $0 \le \beta_j \le \frac{ma}{1+ma}$ and $\sum_j \beta_j = \frac{n(z)a}{1+ma}$, where these constraints follow immediately from $0 \le \widetilde{m}_j \le m$ and $\sum_j \widetilde{m}_j = n(z)$. It is easy to verify that the function $f(\beta)$ is increasing in β, and therefore, the maximal value of (35) is obtained by setting as many values of β_j as possible to the maximum value $\frac{ma}{1+ma}$, and letting an additional value of β_j equal the remainder (if any). This amounts to setting as many values of $\widetilde{m}_j$ as possible to m, and letting an additional value of $\widetilde{m}_j$ equal the remainder. The corresponding maximum value is

$$\sum_{j=1}^{\lfloor \frac{p}{m} \rfloor} \beta_j f(\beta_j) = \Big\lfloor \frac{n(z)}{m} \Big\rfloor \frac{ma}{1+ma} f\Big(\frac{ma}{1+ma}\Big) + \frac{ra}{1+ma} f\Big(\frac{ra}{1+ma}\Big) \tag{36}$$

$$\le \frac{n(z)}{m}\frac{ma}{1+ma} f\Big(\frac{ma}{1+ma}\Big), \tag{37}$$

where r denotes the remainder value (i.e., the additional value of $\widetilde{m}_j$ mentioned above), and (37) follows by writing $f\big(\frac{ra}{1+ma}\big) \le f\big(\frac{ma}{1+ma}\big)$ using the above-mentioned monotonicity of f.

Roughly speaking, we have argued that given a budget of $n(z)$ nodes to observe, the ones that yield a graph that is "furthest" from the empty graph are those that correspond to $\lfloor \frac{n(z)}{m} \rfloor$ complete m-cliques, with any remainder also concentrated within a single clique. Intuitively, this is because taking measurements from a variety of different cliques yields more independent nodes, thus being closer to the behavior of the empty graph in which all nodes are independent.

Upper bounding the summation on the right-hand side of (34) by the maximum value (37), we obtain

$$D(P_{G(z)}\|Q_{(z)}) \le \frac{n(z)}{2m}\bigg(-\log\bigg(1-\frac{ma}{1+ma}\bigg) - \frac{ma}{1+ma}\bigg). \tag{38}$$

Applying the inequality $-\log(1-\frac{\beta}{1+\beta}) - \frac{\beta}{1+\beta} \le \frac{1}{2}\log(1+\beta^2)$, we can weaken (38) to

$$D(P_{G(z)}\|Q_{(z)}) \le \frac{n(z)}{4m}\log\big(1+(ma)^2\big). \tag{39}$$

We obtain from (39) and (12) that

$$I(G;X^{(i)}|Z^{(i)}) \le \frac{\mathbb{E}[n(Z^{(i)})]}{4m}\log\big(1+(ma)^2\big), \tag{40}$$

and summing over i and again noting that $\sum_{i=1}^N n(Z^{(i)}) \le n$ with probability one, we obtain

$$\sum_{i=1}^N I(G;X^{(i)}|Z^{(i)}) \le \frac{n}{4m}\log\big(1+(ma)^2\big). \tag{41}$$

Substitution into (11) yields the necessary condition

$$n \ge \frac{2pd\log\frac{p}{d}}{\log\Big(1+\big((d+1)\frac{\tau}{1-\tau}\big)^2\Big)}(1-\delta-o(1)), \tag{42}$$

where the numerator arises from (28), and we have set $m = d+1$ and $a = \frac{\tau}{1-\tau}$.

5 Discussion: Average Degree vs. Maximal Degree

The question of whether the maximal degree $d_{\max}$ or average degree d_{avg} dictates the performance of active graphical model selection was raised in [7],[1] where it was suggested that it is the latter in the Gaussian case if τ is bounded away from zero. Our results are proved by considering restricted ensembles for which $d_{\text{avg}} = d_{\max}(1 + o(1))$, and hence it is not immediately clear which is more fundamental. We proceed by discussing the two for both Ising models and Gaussian models.

We first remark that the first terms in each of (8) and (9) do not contain d, and they were proved by considering an ensemble where every node has degree exactly one. Moreover, the third term in (8) is trivially obtained by counting the number of graphs with maximal degree d, without any further restrictions, and it is unclear how to adapt this to gain insight on the role of the average degree. Hence, to provide a distinction between $d_{\max}$ and d_{avg}, we focus only on the second terms in (8) and (9).

For the Ising model, the second term in (8) was obtained by considering $\lfloor \frac{p}{d+1} \rfloor$ cliques of size $d+1$, and considering graphs obtained by subsequently removing a single edge, *cf.*, Section 4.2.2. In the supplementary material, we describe an analogous ensemble in which these cliques have different sizes, and show that the term $e^{\lambda d_{\max}}$ still arises in the resulting sample complexity bound. Intuitively, this is because even if all cliques except the largest are known perfectly and an edge is removed from the largest one, it is still very difficult to identify that edge. Hence, regarding this exponential term (which is the main feature of the bound), it is $d_{\max}$ that dictates the performance here.

For the Gaussian model, the second term in (9) was obtained by considering graphs containing $\lfloor \frac{p}{d+1} \rfloor$ cliques of size $d+1$, *cf.*, Section 4.3.2. In the supplementary material, we provide a natural extension of this ensemble which instead uses cliques of *differing* sizes $(d_1, \dots, d_K)$ such that $\sum_{k=1}^{K}(d_k+1) = p$. We make the mild assumption that each of these degrees behaves as $d_k = o(p)$.

The most straightforward extension of the proof of Theorem 2 yields a bound of the form $n = \Omega\big(\frac{p d_{\min} \log \frac{p}{d_{\max}}}{\log(1+\tau d_{\max})}\big)$, where $d_{\min}$ is the minimum degree. This bound is rarely tight, but it can be improved by a genie argument: Reveal to the decoder all of the smallest cliques, up to a total of $(1-\alpha)p$ nodes for some $\alpha \in (0,1)$. The decoder is left to estimate the remaining cliques among αp nodes.

In the supplementary material, we show that as long as α is bounded away from zero and one, this approach yields a sample complexity lower bound of the form $n = \Omega\big(\frac{p d^{(\alpha)}_{\min} \log \frac{p}{d_{\max}}}{\log(1+\tau d_{\max})}\big)$, where $d^{(\alpha)}_{\min}$ is the minimum degree among the remaining αp nodes. If the top αp node degrees in the graph coincide to within a constant factor, then we have $d^{(\alpha)}_{\max} = \Theta(d_{\text{avg}})$, and we thus match the $O((1+d_{\text{avg}})p \log p)$ upper bound from [7] for fixed τ, up to a logarithmic factor.

These observations support the idea proposed in [7] that the average degree is the more fundamental quantity in the Gaussian setting with fixed τ. Note, however, that the assumptions are slightly different, due to the coherence assumption made in [7] and the above assumption on the top αp node degrees.

6 Conclusion

We have provided lower bounds on active learning for graphical model selection. Using a variety of restricted graph ensembles, we recovered analogous bounds to those for the passive setting, suggesting that active learning does not help much in the minimax sense for the degree-bounded class $\mathcal{G}_d$. Moreover, we identified an ensemble for the Ising model in which the maximal degree remains the crucial quantity, and another ensemble for the Gaussian model in which the average degree is the more important quantity. We note that our analysis also readily extends to the edge-bounded class $\mathcal{G}_k$ in which all graphs have at most k edges, analogously to previous works such as [8, 23].

An important direction for further research is to characterize the gain (if any) that can be achieved by active learning in the case of *random* graphs (e.g., Erdös-Rényi [20, 25], power law [21]), in which the maximal and average degrees can differ considerably. Moreover, it would be of interest to understand the role of active learning when the edges have differing parameters $\{\lambda_{ij}\}$ in the Ising model, or when the values $\tau_{ij} = \frac{|\Theta_{ij}|}{\sqrt{\Theta_{ii}\Theta_{jj}}}$ differ in the Gaussian model.

Acknowledgment

This work was supported in part by the European Commission under Grant ERC Future Proof, SNF 200021-146750 and SNF CRSII2-147633, and by the 'EPFL Fellows' programme (Horizon2020 grant 665667).

[1] More precisely, [7] considers the quantity $\overline{d}_{\max}$ defined in Section 1.2, but this coincides with d_{avg} for all ensembles considered in this paper, at least up to a multiplicative $1 + o(1)$ term.

References

[1] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Patt. Analysis and Mach. Intel.*, no. 6, pp. 721–741, 1984.

[2] R. J. Glauber, "Time-dependent statistics of the Ising model," *J. Math. Phys.*, vol. 4, no. 2, pp. 294–307, 1963.

[3] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge Univ. Press, 1998.

[4] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.

[5] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge Univ. Press, 1994, vol. 8.

[6] D. M. Chickering, "Learning Bayesian networks is NP-complete," in *Learning from data*. Springer, 1996, pp. 121–130.

[7] G. Dasarathy, A. Singh, M.-F. Balcan, and J. H. Park, "Active learning algorithms for graphical model selection," in *Int. Conf. Art. Intel. Stats. (AISTATS)*, 2016.

[8] N. Santhanam and M. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4117–4134, July 2012.

[9] W. Wang, M. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *IEEE Int. Symp. Inf. Theory*, 2010.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

[11] E. Arias-Castro, E. J. Candes, and M. A. Davenport, "On the fundamental limits of adaptive sensing," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 472–481, Jan. 2013.

[12] J. Scarlett and V. Cevher, "Phase transitions in group testing," in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2016.

[13] L. Baldassini, O. Johnson, and M. Aldridge, "The capacity of adaptive group testing," in *IEEE Int. Symp. Inf. Theory*, July 2013, pp. 2676–2680.

[14] R. M. Castro and R. D. Nowak, "Minimax bounds for active learning," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2339–2353, 2008.

[15] J. Haupt, R. M. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6222–6235, 2011.

[16] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, 2011.

[17] K. P. Murphy, "Active learning of causal Bayes net structure," 2001, technical report, UC Berkeley.

[18] S. Tong and D. Koller, "Active learning for structure in Bayesian networks," in *Int. Joint Conf. Art. Intel.*, 2001.

[19] G. Bresler, E. Mossel, and A. Sly, "Reconstruction of Markov random fields from samples: Some observations and algorithms," in *Appr., Rand. and Comb. Opt. Algorithms and Techniques*. Springer Berlin Heidelberg, 2008, pp. 343–356.

[20] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky, "High-dimensional structure estimation in Ising models: Local separation criterion," *Ann. Stats.*, vol. 40, no. 3, pp. 1346–1375, 2012.

[21] R. Tandon and P. Ravikumar, "On the difficulty of learning power law graphical models," in *IEEE Int. Symp. Inf. Theory*, 2013.

[22] K. Shanmugam, R. Tandon, A. Dimakis, and P. Ravikumar, "On the information theoretic limits of learning Ising models," in *Adv. Neur. Inf. Proc. Sys. (NIPS)*, 2014.

[23] J. Scarlett and V. Cevher, "On the difficulty of selecting Ising models with approximate recovery," 2016, accepted to *IEEE Trans. Sig. Inf. Proc. over Networks*.

[24] D. Vats and J. M. Moura, "Necessary conditions for consistent set-based graphical model selection," in *IEEE Int. Symp. Inf. Theory*, 2011, pp. 303–307.

[25] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky, "High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion," *J. Mach. Learn. Res.*, vol. 13, pp. 2293–2337, 2012.

[26] V. Jog and P.-L. Loh, "On model misspecification and KL separation for Gaussian graphical models," in *IEEE Int. Symp. Inf. Theory*, 2015.

[27] R. Wu, R. Srikant, and J. Ni, "Learning loosely connected Markov random fields," *Stoch. Sys.*, vol. 3, no. 2, pp. 362–404, 2013.

[28] A. Jalali, C. C. Johnson, and P. K. Ravikumar, "On learning discrete graphical models using greedy methods," in *Adv. Neur. Inf. Proc. Sys. (NIPS)*, 2011.

[29] A. Ray, S. Sanghavi, and S. Shakkottai, "Greedy learning of graphical models with small girth," in *Allteron Conf. Comm., Control, and Comp.*, 2012.

[30] G. Bresler, D. Gamarnik, and D. Shah, "Structure learning of antiferromagnetic Ising models," in *Adv. Neur. Inf. Proc. Sys. (NIPS)*, 2014.

[31] G. Bresler, "Efficiently learning Ising models on arbitrary graphs," in *ACM Symp. Theory Comp. (STOC)*, 2015.

[32] P. Ravikumar, M. J. Wainwright, J. D. Lafferty, and B. Yu, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Ann. Stats.*, vol. 38, no. 3, pp. 1287–1319, 2010.

[33] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Elec. J. Stats.*, vol. 5, pp. 935–980, 2011.

[34] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Stats.*, vol. 34, no. 3, pp. 1436–1462, June 2006.

[35] E. Yang, A. C. Lozano, and P. K. Ravikumar, "Elementary estimators for graphical models," in *Adv. Neur. Inf. Proc. Sys. (NIPS)*, 2014, pp. 2159–2167.

[36] E. Ising, "Beitrag zur theorie des ferromagnetismus," *Zeitschrift für Physik A Hadrons and Nuclei*, vol. 31, no. 1, pp. 253–258, 1925.

[37] S. L. Lauritzen, *Graphical models*. Clarendon Press, 1996.