Learning Markov Graphs Up To Edit Distance

Abhik Das, Praneeth Netrapalli, Sujay Sanghavi and Sriram Vishwanath

Department of ECE, The University of Texas at Austin, USA

Abstract

This paper presents a rate distortion approach to Markov graph learning. It provides lower bounds on the number of samples required for any algorithm to learn the Markov graph structure of a probability distribution, up to edit distance. In particular, for both Ising and Gaussian models on p variables with degree at most d, we show that at least $\Omega((d - \frac{s}{p}) \log p)$ samples are required for any algorithm to learn the graph structure up to edit distance s. Our bounds represent a strong converse; i.e., we show that for a lower number of samples, the probability of error goes to 1 as the problem size increases. These results show that substantial gains in sample complexity may not be possible without paying a significant price in edit distance error.

1 Introduction

Markov networks, also known as (undirected) graphical models, describe the interdependencies (or the lack of thereof) among a collection of random variables using an undirected graph. As such, they have been used for modeling and designing applications in a multitude of settings, for example, social network modeling [1], [2], image processing/computer vision [3], [4] and computational biology [5], [6]. The problem of learning the graph structure of a Markov network from samples generated by the underlying graph structure is a well-studied one and is referred to as the problem of graphical model selection. There is diverse literature on various aspects of learning graphical models, from statistical physics to computational learning theory. It is only relatively recently that the information theoretic limits for the problem of graphical model selection are being better understood.

An understanding of the information theoretic limits of high-dimensional learning problems in general, and the problem of graphical model selection in particular, provides us with sample complexity bounds corresponding to lower bounds for learning. A useful tool used for obtaining these bounds is Fano's inequality and its generalizations [7]. However, Fano's inequality results in weak converse bounds – the typical result obtained for graphical model selection problem is that if the number of observed samples available to a learning algorithm falls below a certain threshold, the probability of error is bounded away from zero (for example, exceeds 1/2). Therefore, alternate

information-theoretic tools are required for stronger bounds on sample complexity. This motivates the the formulation of strong converse type results, in the same spirit as in the case of noisy channel coding [8]. In other words, we desire results for the graphical model selection problem that state that unless the number of available samples exceeds some threshold, the probability of error in learning the structure of a Markov network goes to 1 as the problem size increases. Such information-theoretic limits are important as they provide an understanding of settings where recovery is impossible, regardless of the algorithm or the cleverness of its design.

In this paper, we focus on reconstruction of the graph structure of a Markov network within a pre-specified distortion, rather than exact reconstruction. As an overarching goal, we are interested in characterizing the rate-distortion limits of the problem of graphical model selection. We restrict ourselves to Markov networks whose underlying graphical structures have bounded degree. We derive detailed results for two well-known families – Ising models and Gaussian Markov networks. The distortion metric we choose is edit distance, that we define in the next section.

1.1 Related Work

There is a significant body of literature in the context of deriving the information-theoretic limits on the sample complexity for exact learning of Markov networks, especially the specialized cases of Ising models [9], [10], and Gaussian Markov networks [11], [12]. The graph ensembles that have been considered include degree-bounded graphs [9]–[12],[13], graphs with limited edges [9] and random graphs [10], [12]. A common theme in deriving these theoretical bounds is to treat the graphical model selection problem as a noisy channel coding problem and apply Fano's inequality to characterize the limits. The graphical model selection problem in presence of distortion has been examined in [12] for the ensemble of Erdös-Rényi graphs. The only known strong converse results have been derived in [10] and [14], for the cases of exact reconstruction of Ising models based on Erdös-Rényi graphs and Gaussian Markov networks based on degree-bounded graphs respectively. The performance of graphical model learning algorithms, that output a set of graphs instead of a single one (similar to list-decoding), is examined in [15] for Gaussian and Ising models.

1.2 Summary of Results

We provide a comparison of our results in this paper vs. existing ones in literature in Table 1. All the results are for the ensemble of graphs on p nodes with degree at most d. Our results are highlighted in bold face. s denotes the maximum allowed edit distance between the original and recovered graphs. All existing results in literature are for the case of exact recovery, i.e., $s = 0$.

It is known that the edge weights play an important role in determining the complexity of learning graphical models [16]. However this dependency is complex in general. For instance both very low edge weights and very large edge weights increase the sample complexity of learning the

Table 1: Comparison with existing results

Model	Edge weight $= \Theta\left(\frac{1}{d}\right)$	Edge weight $= \Theta\left(\frac{1}{\sqrt{d}}\right)$
Ising	$\Omega(d^2 \log p)$ [9]	$\Omega(\sqrt{d}\exp(\sqrt{d})\log p)$ [9]
	$\boldsymbol{\Omega\left(\left(d - \frac{8s}{p}\right)\log p\right)}$	$\boldsymbol{\Omega\left(\left(d - \frac{8s}{p}\right)\log p\right)}$
Gaussian	$\Omega\left(d^2 \log p\right)$ [11]	$\Omega\left(d \log p\right)$ [11]
	$\boldsymbol{\Omega\left(\sqrt{d}\left(d - \frac{4s}{p}\right)\log p\right)}$	$\boldsymbol{\Omega\left(\left(d - \frac{4s}{p}\right)\log p\right)}$

graphical model - the first because of difficulty in recognizing the presence of an edge and the latter because of large range correlations. Existing results show significant difference in lower bounds for edge weight scaling as $\Theta\left(\frac{1}{d}\right)$ as opposed to $\Theta\left(\frac{1}{\sqrt{d}}\right)$. However, in this paper, we are able to show different lower bounds for the Gaussian case but not for the Ising model. A detailed description as well as comparison of our results with the existing ones is done in Section 6.

The rest of this paper is organized as follows. We discuss some preliminaries and introduce the system model in Section 2. We consider the problem of learning graph structure up to a pre-specified distortion value (edit distance) and give strong limits on the sample complexity for arbitrary ensembles, Ising models and Gaussian Markov networks in Sections 3, 4 and 5 respectively. We finally conclude the paper with Section 6 and present the proofs in the appendix.

2 Preliminaries

We consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes and $E \subset V \times V$ is the set of edges. A Markov network is obtained by associating a random variable X_i to node $i \in V$, that takes values from an alphabet set $\mathcal{A}$, and specifying a joint probability distribution p over vector $X := (X_1, X_2, \ldots, X_p)$ that satisfies the following property:

$$p(x_A, x_B|x_C) = p(x_A|x_C)p(x_B|x_C),$$

where A, B and C are any disjoint subsets of V such that every path in G from A to B passes through C, and x_A, x_B and x_C denote the restrictions of $(x_1, \ldots, x_p) \in \mathcal{A}^p$ to indices in A, B and C respectively. Note that $p(\cdot)$ denotes p.m.f. in the discrete case (i.e., $\mathcal{A}$ is a finite set) and p.d.f. in the continuous case (i.e., $\mathcal{A} = \mathbb{R}$ and the distribution is a continuous function). We briefly describe the families of discrete and continuous graphical models considered in this paper.

Ising Model: This is a discrete probability distribution that is widely studied in statistical physics [17], and applicable to fields like computer vision [18] and game theory [19]. In this paper we consider a special case of Ising model, called the zero field binary Ising model. Here the alphabet is

$\mathcal{A} = \{-1, 1\}$. Given an undirected graph G on p nodes and weight $\theta_{ij} \in \mathbb{R}$ for each edge $(i,j) \in E$, the probability of a configuration $x = (x_1, x_2, \ldots, x_p) \in \mathcal{A}^p$ is given by the following expression:

$$p\left(x\right) = \frac{\exp\left(\sum_{(i,j)\in E} \theta_{ij} x_i x_j\right)}{\sum_{x\in\mathcal{A}^p} \exp\left(\sum_{(i,j)\in E} \theta_{ij} x_i x_j\right)}.$$

Gaussian Graphical Model: This is one of the most well known families of continuous Markov networks. Here X has a multivariate Gaussian distribution. Without any loss of generality, it can be assumed that X has zero mean. Given an undirected graph G on p nodes and a $p \times p$ positive definite matrix Θ, the p.d.f. of X is given by the following expression:

$$p(x) = \frac{1}{\sqrt{(2\pi)^p |\Theta^{-1}|}} \exp\left(-\frac{1}{2} x^T \Theta x\right),$$

where $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$. Θ is the inverse covariance matrix and is such that $\Theta(i,j) \neq 0$ if and only if $(i,j) \in E$. Θ is also called the potential matrix, since $\Theta(i,j)$ can be interpreted as the potential of edge $(i,j) \in E$. The following quantity, called minimum magnitude of partial corelation coefficient, plays an important role in determining the complexity of the structure learning problem:

$$\lambda^*(\Theta) := \min_{(i,j)\in E} \frac{|\Theta(i,j)|}{\sqrt{\Theta(i,i)\Theta(j,j)}}.$$

This quantity is invariant to rescaling of the random variables and can be thought of as the minimum magnitude of a non-zero entry after normalizing the diagonal terms of Θ to one.

In this paper we restrict our attention to the ensemble of degree bounded graphs in light of the fact that extensive work on learning graphical models focus on these graphs. We denote the set of all graphs on p nodes and maximum degree d by $\mathcal{G}_{p,d}$. We also denote the set of all graphs on p nodes by $\mathcal{U}_p$. For any two graphs G and H on the same set of nodes, we define the *edit distance* $\Delta(G,H)$ as the minimum number of edge deletions and insertions required to convert G to H. In other words, $\Delta(G,H)$ is the cardinality of the symmetric difference between edge sets of G and H.

2.1 Learning Algorithm and Error Criterion

We consider an ensemble of undirected graphs on a common set of p nodes, $\mathcal{G} = \{G_1, \ldots, G_M\}$, and an ensemble of Markov networks $\mathcal{K} = \{K_1, \ldots, K_M\}$, such that K_i is associated with G_i and the random variables $X = (X_1, \ldots, X_p)$ draw values from alphabet $\mathcal{A}$. We choose a Markov network $K \in \mathcal{K}$ uniformly at random and obtain n i.i.d. vector samples $X^n = (X^{(1)}, \ldots, X^{(n)})$ from the distribution specified by K. The problem we consider is to reconstruct the graph G associated with K given the samples X^n. This is also known as the problem of graphical model selection. A

function $\phi : \mathcal{A}^{np} \to \mathcal{U}_p$ that maps the observed samples to an estimated graph $\hat{G} = \phi(X^n) \in \mathcal{U}_p$ is called a learning algorithm. Given a pre-specified s, we define the error event for the learning algorithm as $\{\Delta(G, \phi(X^n)) \geq s\}$, i.e., the algorithm is correct only if the edit distance between the actual and estimated graphs is less than s. Then the probability of error can be defined as

$$P_e^{(n)}(\phi) = P(\Delta(G, \phi(X^n)) \geq s) = \frac{1}{M} \sum_{i=1}^{M} P\left(\Delta(G_i, \phi(X^n)) \geq s | K = K_i\right).$$

In this paper, we derive lower bounds on the sample size n, in terms of the ensemble parameters, for any learning algorithm to reliably recover the underlying graph of a Markov network up to an edit distance of s. We do this by bounding $P_e^{(n)}(\phi)$ in terms of n and the ensemble parameters.

3 Lower Bounds for Arbitrary Ensembles

In this section, we state our result for lower bounds on the sample complexity for arbitrary ensembles of graphical models. We consider the same setup as described in Section 2.1. We have an ensemble of Markov network models $\mathcal{K}$ and the corresponding ensemble of undirected graphs $\mathcal{G}$ on p nodes. We choose a Markov network $K \in \mathcal{K}$ uniformly at random and obtain n i.i.d. sample vectors from its joint distribution. Our aim is to analyze the performance of *any* learning algorithm $\phi : \mathcal{A}^{np} \to \mathcal{U}_p$ and bound its probability of error. For this we define the following quantities for $G \in \mathcal{G}$:

$$\begin{aligned} B(s, G) &= \{H : \Delta(G, H) < s, H \in \mathcal{U}_p\}, \\ B(s, \mathcal{G}) &= \max_{G \in \mathcal{G}} |B(s, G)|. \end{aligned}$$

$B(s, \mathcal{G})$ represents the maximum number of graphs in $\mathcal{U}_p$ that are at an edit distance of at most s from any graph in $\mathcal{G}$. We also define another quantity, similar in structure to mutual information:

$$I(K_i; X^{(1)}) = \begin{cases} H(X^{(1)}) - H(X^{(1)}|K = K_i) & \text{if } |\mathcal{A}| < \infty, \\ h(X^{(1)}) - h(X^{(1)}|K = K_i) & \text{if } \mathcal{A} = \mathbb{R}. \end{cases}$$

We define a lower bound R and an upper bound C on the following quantities:

$$R \leq \log M - \log B(s, \mathcal{G}), \tag{1}$$

$$C \geq \max_{1 \leq i \leq M} I(K_i; X^{(1)}). \tag{2}$$

Then we have the following theorem which establishes a necessary condition on the number of samples n for consistent recovery of the graphical model using any learning algorithm.

Theorem 1. *Consider an ensemble of Markov networks $\mathcal{K} = \{K_1, \ldots, K_M\}$ and the corresponding ensemble of undirected graphs on p nodes, $\mathcal{G} = \{G_1, \ldots, G_M\}$. Suppose the random variables take values from alphabet $\mathcal{A}$. If the number of samples satisfies $n < \frac{R}{C}$ then for any learning algorithm, we have the following lower bound on the probability of error:*

$$P_e^{(n)}(\phi) \geq 1 - \frac{4nA(\mathcal{K})}{(R-nC)^2} - 2^{-\frac{(R-nC)}{2}}.$$

Here error means that the edit distance between the original and recovered graphs exceeds s and

$$A(\mathcal{K}) = \max_{1 \leq i \leq M} var\left(\log \frac{p\left(X^{(1)} \middle| K = K_i\right)}{p(X^{(1)})} \middle| K = K_i \right).$$

where $p(\cdot)$ stands for p.m.f. in the discrete case and p.d.f. in the continuous case.

If we find a good upper bound for $A(\mathcal{K})$ for a given ensemble, we can use Theorem 1 to show that $P_e^{(n)} \to 1$ in the high dimensional setting as $p \to \infty$ and $n < \frac{R}{C}$. We pursue this approach in the next two sections to prove results for ensembles of Ising and Gaussian graphical models.

4 Lower Bounds for Ising Models

Our main result in this section is the following theorem which characterizes a lower bound on the number of samples required for consistent recovery of graphical model from an ensemble $\mathcal{K}_{p,d}^I$, whose construction is described below. For each graph $G \in \mathcal{G}_{p,d}$, consider the corresponding Ising model with all edge parameters equal to θ where $\theta \in \left(0, \frac{1}{\sqrt{d}}\right)$. We denote this ensemble of Ising models by $\mathcal{K}_{p,d}^I$. Note that there is a bijection between graphs in $\mathcal{G}_{p,d}$ and models in $\mathcal{K}_{p,d}^I$.

Theorem 2. *Suppose K is chosen uniformly at random from $\mathcal{K}_{p,d}^I$. If $d = o(p^\alpha)$ for some $0 \leq \alpha < 1$, $s < \frac{(1-\alpha)}{16} pd$ and the number of samples n, available to a learning algorithm, satisfies*

$$n < \frac{1}{2}\left[\left(\frac{d}{4} - \frac{2s}{p}\right)\log p - \frac{d}{4}\log 8d + \frac{s}{p}\log\frac{2s}{e} - \frac{\log s}{p}\right] = \Omega\left(\left((1-\alpha)d - \frac{8s}{p}\right)\log p\right),$$

then for any graphical model learning algorithm, the probability of error $P_e^{(n)} \to 1$ as $p \to \infty$.

Proof strategy for Theorem 2: The proof of Theorem 2 follows from establishing the bounds R and C in (1) and (2) and then using Theorem 1. Lemmas 1 and 2 establish such bounds R and C respectively. A complete proof of Theorem 2 can be found in the appendix. We list the graphs in $\mathcal{G}_{p,d}$ as $\{G_1, \ldots, G_M\}$ and the corresponding Ising models in $\mathcal{K}_{p,d}^I$ as $\{K_1, \ldots, K_M\}$. Now we state the following two lemmas bounding R and C for this ensemble.

Lemma 1. *For graph ensemble $\mathcal{G}_{p,d}$ with $d \leq \frac{(p-1)}{2}$ and $s \leq \frac{p(p-1)}{4}$, the following bounds hold:*

$$\log M \geq \frac{pd}{4} \log \frac{p}{8d}, \quad B(s, \mathcal{G}_{p,d}) < s\binom{p^2/2}{s}.$$

Lemma 2. *Suppose K is chosen uniformly at random from $\mathcal{K}^I_{p,d}$. Then we have*

$$\max_i I(K_i; X^{(1)}) \leq p.$$

5 Lower Bounds for Gaussian Markov Graphs

Our main result in this section is a lower bound on the number of samples required for consistent recovery of graphical model from an ensemble of Gaussian Markov networks $\mathcal{K}^G_{p,d}$, which is constructed as follows. Without loss of generality, we assume that p is even. We choose d perfect matchings on p nodes, each perfect matching chosen uniformly at random, and form a multigraph resulting from the union of the matchings. We refer to the set of all such multigraphs on p nodes, constructed in this fashion, as $\mathcal{H}$. The uniform distribution on the set of perfect matchings also defines a probability distribution over $\mathcal{H}$. We have the following lemma for this distribution [20]:

Lemma 3. *Consider a multigraph $H = (V, E)$, $V = \{1, 2, \ldots, p\}$, formed from the union of d random perfect matchings on V, the matchings being chosen according to a uniform distribution. Suppose the eigenvalues of the (weighted) adjacency matrix of H, denoted by A, are $d = \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_p(A)$. Define $\rho(A) = \max_{2 \leq i \leq p} |\lambda_i(A)|$. Then the following result holds:*

$$P(\rho(A) < 3\sqrt{d}) \geq 1 - \frac{c}{p^\tau},$$

where c is a positive real and $\tau = \left\lceil \frac{\sqrt{d-1}+1}{2} \right\rceil - 1$.

Next, we eliminate the multigraphs from $\mathcal{H}$ whose (weighted) adjacency matrices A satisfy $\rho(A) \geq 3\sqrt{d}$ and get a reduced set $\mathcal{H}'$. By Lemma 3, $\mathcal{H} \backslash \mathcal{H}'$ forms a small fraction of $\mathcal{H}$ as $p \to \infty$. We fix constants $\lambda \in \left(0, \frac{1}{4\sqrt{d}}\right)$, $\delta > 0$ and define $\mu := \frac{\delta}{\lambda^{-1} - 4\sqrt{d}}$. Then for every multigraph $H \in \mathcal{H}'$, we generate a $p \times p$ matrix $\Theta = (4\sqrt{d}\mu + \delta)I_p + \mu A$, where I_p is the $p \times p$ identity matrix and A is the (weighted) adjacency matrix of multigraph H. We refer to the resulting set of these matrices as $\mathcal{T}$. Then the following property holds for this set:

Lemma 4. *The matrices in $\mathcal{T}$ are symmetric and positive definite.*

Proof. By construction, any matrix $\Theta \in \mathcal{T}$ has the form $\Theta = (4\sqrt{d}\mu + \delta)I_p + \mu A$, where A is the (weighted) adjacency matrix of some multigraph $H \in \mathcal{H}'$, which makes it symmetric. Also, the construction of $\mathcal{H}'$ ensures that $\rho(A) < 3\sqrt{d}$. Therefore, the minimum eigenvalue of Θ is at least

$4\sqrt{d}\mu + \delta - \rho(A)\mu > \sqrt{d}\mu + \delta > 0$. This and the symmetry of Θ ensure that all the eigenvalues of Θ are positive. Hence $\Theta \in \mathcal{T}$ is a symmetric and positive definite matrix. □

Note that the choice of μ ensures that $\lambda^*(\Theta) = \lambda$ for $\Theta \in \mathcal{T}$. Lemma 4 suggests that the matrices of $\mathcal{T}$ can be the inverse covariance matrices of Gaussian Markov networks. By construction, the underlying graph of each of these Markov networks comes from $\mathcal{G}_{p,d}$. We denote this ensemble of Gaussian Markov networks by $\mathcal{K}^G_{p,d}$ and the corresponding graph ensemble by $\mathcal{G}'_{p,d} \subseteq \mathcal{G}_{p,d}$.

Theorem 3. *Suppose K is chosen uniformly at random from $\mathcal{K}^G_{p,d}$. If $d = o(p^\alpha)$ for some $0 \leq \alpha < \frac{1}{2}$, $s < \frac{(1-2\alpha)}{8}pd$ and the number of samples n, available to a learning algorithm, satisfies*

$$n < \frac{\left(d - \frac{4s}{p}\right)\log p - 2d\log 2d + \frac{2s}{p}\log\frac{2s}{e} - \frac{2\log s}{p} - \frac{2}{p}}{2\log\left(1 + \frac{4\sqrt{d}}{\lambda^{-1} - 4\sqrt{d}}\right)} = \Omega\left(\frac{\left((1-2\alpha)d - \frac{4s}{p}\right)\log p}{\log\left(1 + \frac{4\sqrt{d}}{\lambda^{-1} - 4\sqrt{d}}\right)}\right),$$

then for any graphical model learning algorithm, the probability of error $P_e^{(n)} \to 1$ as $p \to \infty$.

Proof strategy for Theorem 3: Analogous to the proof of Theorem 2, the proof of Theorem 3 follows from establishing the bounds R and C in (1) and (2) and then using Theorem 1. Lemmas 5 and 6 establish such bounds R and C respectively. A complete proof of Theorem 3 can be found in the appendix. We list the graphs in $\mathcal{G}'_{p,d}$ as $\{G_1, \ldots, G_M\}$ and the Gaussian models in $\mathcal{K}^G_{p,d}$ as $\{K_1, \ldots, K_M\}$. Now we state the following two lemmas bounding R and C for this ensemble.

Lemma 5. *For graph ensemble $\mathcal{G}'_{p,d}$, the following bounds hold for large enough p:*

$$\log M \geq \frac{pd}{2}\log\frac{p}{4d^2} - 1, \quad B(s, \mathcal{G}'_{p,d}) < s\binom{p^2/2}{s}.$$

Lemma 6. *Suppose K is chosen uniformly at random from $K^G_{p,d}$. Then we have*

$$\max_i I(K_i; X^{(1)}) \leq \frac{p}{2}\log\left(1 + \frac{4\sqrt{d}}{\lambda^{-1} - 4\sqrt{d}}\right).$$

6 Conclusion & Discussion

Remarks about Theorems 2 & 3: Specializing our result to the case of exact recovery, i.e., setting $s = 0$ yields weaker lower bounds on sample complexity than existing results. However, for the Gaussian case [11] with edge weights $\Theta\left(\frac{1}{\sqrt{d}}\right)$ our result matches the existing result, and for both the Ising [9] and Gaussian [11] cases with edge weights $\Theta\left(\frac{1}{d}\right)$, our result is only a factor of d and $\sqrt{d}$ away respectively from existing results. This gap is either due to a limitation of our proof

technique or due to the difference in the kind of guarantee. Specifically, the lower bound results in [9] and [11] use Fano's inequality to obtain a weak converse i.e., if the number of samples n scales below a certain threshold then the probability of error is lower-bounded by a constant as $p \to \infty$. On the other hand our result establishes a strong converse i.e., if the number of samples n scales below a certain threshold then as $p \to \infty$ the probability of error converges to 1.

In this paper, we develop a rate-distortion framework for graph learning, where we characterize lower bounds on sample complexity within a given distortion criterion. We use a strong converse framework to derive these bounds, indicating that it is near-impossible to learn the graphical model with any fewer samples. Our results show that, for both Ising and Gaussian models on p variables with maximum degree d, at least $\Omega\left(\left(d - \frac{s}{p}\right)\log p\right)$ samples are required for any learning algorithm to recover the graph structure to within edit distance s.

References

[1] A. Grabowski and R. Kosinski, "Ising-based model of opinion formation in a complex network of interpersonal interactions," *Physica A: Statistical Mechanics and its Applications*, vol. 361, pp. 651–664, 2006.

[2] F. Vega-Redondo, *Complex social networks*. Cambridge Press, 2007.

[3] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society Series B*, vol. 48, pp. 259–279, 1986.

[4] M. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchial context on a large database of object categories," in *IEEE CVPR*, 2010.

[5] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, Feb 2004.

[6] A. Ahmedy, L. Song, and E. P. Xing, "Time-varying networks: Recovering temporally rewiring genetic networks during the life cycle of drosophila melanogaster," tech. rep., 2008. arXiv.

[7] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley Interscience, 2006.

[8] R. Gallager, *Information Theory and Reliable Communication*. Wiley, 1968.

[9] N. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *arXiv*, 2009.

[10] A. Anandkumar, V. Y. F. Tan, and A. Willsky, "High-dimensional structure learning of Ising models: Tractable graph families," *arXiv Preprint*, 2011.

[11] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *IEEE ISIT*, 2010.

[12] A. Anandkumar, V. Y. F. Tan, and A. Willsky, "High-dimensional Gaussian graphical model selection: Tractable graph families," *arXiv Preprint*, 2011.

[13] G. Bresler, E. Mossel, and A. Sly, "Reconstruction of Markov random fields from samples: Some observations and algorithms," in *APPROX*, pp. 343–356, 2008.

[14] I. Mitliagkas and S. Vishwanath, "Strong information-theoretic limits for source/model recovery," in *Proc. of Allerton Conf. on Communication, Control and Computing, Monticello, USA*, 2010.

[15] D. Vats and J. Moura, "Necessary conditions for consistent set-based graphical model selection," in *IEEE ISIT*, 2011.

[16] A. Montanari and J. A. Pereira, "Which graphical models are difficult to learn?," in *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 1303–1311, 2009.

[17] L. Reichl and J. Luscombe, "A modern course in statistical physics," *American Journal of Physics*, vol. 67, p. 1285, 1999.

[18] S. Geman and C. Graffigne, "Markov random field image models and their applications to computer vision," in *Proceedings of the International Congress of Mathematicians*, vol. 1, p. 2, AMS, Providence, RI, 1986.

[19] Y. Zhang, "Modeling market mechanism with evolutionary games," *Arxiv preprint cond-mat/9803308*, 1998.

[20] J. Friedman, "A proof of Alon's second eigenvalue conjecture and related problems," *arXiv*, 2004.

A Proofs

Proof (Theorem 1). The proof technique is similar to the standard strong converse proofs in information theory [8]. We prove the theorem for the case where $\mathcal{A}$ is a finite set and $\mathcal{K}$ is an ensemble of discrete Markov networks. The proof for the case when $\mathcal{A} = \mathbb{R}$ and $\mathcal{K}$ is an ensemble of Markov networks with continuous density functions goes along the same line.

We fix $\epsilon > 0$ and define the following sets:

$$\mathcal{B}_i = \left\{ x^n \in \mathcal{A}^{np} : \log \frac{p(x^n|K = K_i)}{p(x^n)} \geq n(C + \epsilon) \right\},\ i = 1, 2, \ldots, M.$$

The set $\mathcal{B}_i$ tries to capture all points in the sample space where the random variable $\log \frac{p(X^n|K=K_i)}{p(X^n)}$ is greater than its mean conditioned on $K = K_i$ (strictly speaking, it only contains a subset of those points since C is an upper bound on the mean of the random variable).

Given a learning algorithm $\phi : \mathcal{A}^{np} \to \mathcal{U}_p$, we define the following sets:

$$\mathcal{R}_i = \{x^n \in \mathcal{A}^{np} : \Delta(\phi(x^n), G_i) < s\},\ i = 1, 2, \ldots, M,$$
$$\mathcal{S}_i = \{x^n \in \mathcal{A}^{np} : \phi(x^n) = G_i\},\ i = 1, 2, \ldots, M.$$

$\mathcal{R}_i$ is the set of points in the sample space for which we have correct recovery if $K = K_i$ and $\mathcal{S}_i$ is the set of points in the sample space for which the learning algorithm returns G_i. Enumerating the graphs in $B(s, G_i)$ as $G_{i_1}, \ldots, G_{i_k}$ we see that $\mathcal{R}_i = \cup_{t=1}^{k} \mathcal{S}_{i_t}$. Note that $\mathcal{S}_i \cap \mathcal{S}_j = \phi$ for $i \neq j$. The probability of correct decoding by the learning algorithm is given by

$$\begin{aligned} P_c^{(n)} &= \frac{1}{M} \sum_{i=1}^{M} \sum_{x^n \in \mathcal{R}_i} p(x^n|K = K_i) \\ &= \frac{1}{M} \sum_{i=1}^{M} \sum_{x^n \in \mathcal{R}_i \cap \mathcal{B}_i^c} p(x^n|K = K_i) + \frac{1}{M} \sum_{i=1}^{M} \sum_{x^n \in \mathcal{R}_i \cap \mathcal{B}_i} p(x^n|K = K_i) \end{aligned}$$

The first term can be bounded as follows:

$$\begin{aligned} \frac{1}{M} \sum_{i=1}^{M} \sum_{x^n \in \mathcal{R}_i \cap \mathcal{B}_i^c} p(x^n|K = K_i) &\leq \frac{1}{M} \sum_{i=1}^{M} \sum_{x^n \in \mathcal{R}_i \cap \mathcal{B}_i^c} 2^{n(C+\epsilon)} p(x^n) \\ &= \frac{2^{n(C+\epsilon)}}{M} \sum_{i=1}^{M} \sum_{x^n \in \mathcal{R}_i \cap \mathcal{B}_i^c} p(x^n) \\ &\leq \frac{2^{n(C+\epsilon)}}{M} \sum_{i=1}^{M} \sum_{x^n \in \mathcal{R}_i} p(x^n) \\ &\leq \frac{2^{n(C+\epsilon)}}{M} \sum_{i=1}^{M} \sum_{t=1}^{|B(s,G_i)|} \sum_{x^n \in \mathcal{S}_{i_t}} p(x^n) \\ &\leq \frac{2^{n(C+\epsilon)}}{M} |B(s, \mathcal{G})| \\ &\leq 2^{n(C+\epsilon)-R}. \end{aligned}$$

The second term can be bounded as follows:

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}\sum_{x^n\in\mathcal{R}_i\cap\mathcal{B}_i} p(x^n|K=K_i) &\le \frac{1}{M}\sum_{i=1}^{M}\sum_{x^n\in\mathcal{B}_i} p(x^n|K=K_i)\\ &= \frac{1}{M}\sum_{i=1}^{M} P\left(\log\frac{p(X^n|K=K_i)}{p(X^n)} \ge n(C+\epsilon)\,\middle|\, K=K_i\right)\\ &= \frac{1}{M}\sum_{i=1}^{M} P\left(\sum_{j=1}^{n}\log\frac{p(X^{(j)}|K=K_i)}{p(X^{(j)})} \ge n(C+\epsilon)\,\middle|\, K=K_i\right).\end{aligned}$$

Since $X^{(1)},\ldots,X^{(n)}$ are i.i.d. vectors, we have the following equality:

$$\operatorname{var}\left(\sum_{j=1}^{n}\log\frac{p(X^{(j)}|K=K_i)}{p(X^{(j)})}\,\middle|\, K=K_i\right) = n\operatorname{var}\left(\log\frac{p(X^{(1)}|K=K_i)}{p(X^{(1)})}\,\middle|\, K=K_i\right).$$

where $\operatorname{var}(\cdot)$ is the variance of the random variable. Defining

$$A(\mathcal{K}) = \max_{1\le i\le M}\operatorname{var}\left(\log\frac{p(X^{(1)}|K=K_i)}{p(X^{(1)})}\,\middle|\, K=K_i\right),$$

and using (2) along with Chebyshev's inequality, we obtain the following bound:

$$P\left(\sum_{j=1}^{n}\log\frac{p(X^{(j)}|K=K_i)}{p(X^{(j)})} \ge n(C+\epsilon)\,\middle|\, K=K_i\right) \le \frac{A(\mathcal{K})}{n\epsilon^2},\quad i=1,2,\ldots,M.$$

Choosing $\epsilon=(R-nC)/2n$ and the fact that $P_e^{(n)}=1-P_c^{(n)}$ gives us

$$P_e^{(n)} \ge 1-\frac{4nA(\mathcal{K})}{(R-nC)^2}-2^{-\frac{(R-nC)}{2}}.$$

□

Proof (Lemma 1). The first inequality follows from the proof of Theorem 1 in [9]. For the second inequality, note that for graphs $G=(V,E(G))\in\mathcal{G}_{p,d}$ and $H=(V,E(H))\in\mathcal{U}_p$ with $\Delta(G,H)<s$, we have $|E(G,H)|<s$, where $E(G,H)$ is the symmetric difference of the edge sets $E(G)$ and $E(H)$. In other words, $(V,E(G,H))$ is a graph on p nodes and at most $s-1$ edges. Therefore, $B(s,\mathcal{G}_{p,d})$

is no more than the number of graphs on p nodes and at most $s-1$ edges. This gives

$$B(s, \mathcal{G}_{p,d}) \leq \sum_{i=0}^{s-1} \binom{p(p-1)/2}{i} \leq s\binom{p(p-1)/2}{s} < s\binom{p^2/2}{s},$$

where we use the fact that $\binom{m}{i} \leq \binom{m}{j}$ for $0 \leq i \leq j \leq m/2$ and $s \leq p(p-1)/4$. □

Proof (Lemma 2). Since $X^{(1)}$ is a vector of p random variables taking values from $\{-1, 1\}$, we have $H(X^{(1)}) \leq p$. Hence, by definition, $\max_i I(K_i; X^{(1)}) \leq H(X^{(1)}) \leq p$. □

Proof (Theorem 2). Using Lemmas 1 and 2, we can define bounds R and C in (1) and (2) as follows:

$$R := \frac{pd}{4} \log \frac{p}{8d} - \log\left(s\binom{p^2/2}{s}\right), \tag{3}$$

$$C := p. \tag{4}$$

Using the fact that $\binom{a}{b} \leq \left(\frac{a \cdot e}{b}\right)^b$, we obtain the following lower bound on $\frac{R}{C}$:

$$\frac{R}{C} \geq \left(\frac{d}{4} - \frac{2s}{p}\right) \log p - \frac{d}{4} \log 8d + \frac{s}{p} \log \frac{2s}{e} - \frac{\log s}{p}.$$

Under the hypothesis of the theorem, we have $n < \frac{R}{2C}$. Theorem 1 shows that $P_e^{(n)}$ satisfies

$$P_e^{(n)} \geq 1 - \frac{8A(\mathcal{K}_{p,d}^I)}{RC} - 2^{-\frac{R}{4}}. \tag{5}$$

We need to now show that the last two terms of the RHS of (5) go to 0 as $p \to \infty$. Since $d = o(p^\alpha)$ for some $\alpha < 1$ and $s < \frac{(1-\alpha)pd}{16}$, we have

$$R = \Theta(pd \log p). \tag{6}$$

This shows that the last term of the RHS of (5) goes to 0 as $p \to \infty$. To show the same for the second last term, we give a bound for $A(\mathcal{K}_{p,d}^I)$. For this, we recall the definition of $A(\mathcal{K}_{p,d}^I)$:

$$\begin{aligned} A(\mathcal{K}_{p,d}^I) &= \max_{1 \leq i \leq M} \operatorname{var}\left(\log \frac{p\left(X^{(1)} \middle| K = K_i\right)}{p(X^{(1)})} \middle| K = K_i\right) \\ &\leq \max_{1 \leq i \leq M} \mathbb{E}\left[\left(\log \frac{p\left(X^{(1)} \middle| K = K_i\right)}{p(X^{(1)})}\right)^2 \middle| K = K_i\right]. \end{aligned} \tag{7}$$

To bound $A(\mathcal{K}_{p,d}^I)$, we give a deterministic bound on $\log \frac{p\left(X^{(1)} \middle| K=K_i\right)}{p(X^{(1)})}$ and use (7). Note that for

$K_i \in \mathcal{K}^I_{p,d}$, the total number of edges in the corresponding graph G_i does not exceed $\frac{pd}{2}$. Also, for every edge (k,l) in E_i, the edge set of G_i, we have $|\theta_{kl}| = \theta = O\left(\frac{1}{\sqrt{d}}\right)$. Then given $K_i \in \mathcal{K}^I_{p,d}$ and $x \in \{-1,1\}^p$, we have the following upper bound:

$$p\left(x|K = K_i\right) = \frac{\exp\left(\sum_{(k,l)\in E_i} \theta_{kl} x_k x_l\right)}{\sum_{x\in\{-1,1\}^p} \exp\left(\sum_{(k,l)\in E^i} \theta_{kl} x_k x_l\right)} \leq \frac{\exp\left(\theta\frac{pd}{2}\right)}{2^p \exp\left(-\theta\frac{pd}{2}\right)} = \exp\left(p\left(\theta d - \log 2\right)\right). \quad (8)$$

Similarly we have the following lower bound :

$$p\left(x|K = K_i\right) = \frac{\exp\left(\sum_{(k,l)\in E_i} \theta_{kl} x_k x_l\right)}{\sum_{x\in\{-1,1\}^p} \exp\left(\sum_{(k,l)\in E^i} \theta_{kl} x_k x_l\right)} \geq \frac{\exp\left(-\theta\frac{pd}{2}\right)}{2^p \exp\left(\theta\frac{pd}{2}\right)} = \exp\left(-p\left(\theta d + \log 2\right)\right). \quad (9)$$

Since $p(x)$ is the average of $p\left(x|K = K_i\right)$, we have the same bounds as above for $p(x)$. Using this observation as well as bounds (8) and (9) we obtain the following inequality:

$$\left|\log \frac{p(x|K = K_i)}{p(x)}\right| \leq \left|\log \frac{\exp\left(p\left(\theta d - \log 2\right)\right)}{\exp\left(-p\left(\theta d + \log 2\right)\right)}\right| = 2p\theta d \log e. \quad (10)$$

Since (10) holds for $i = 1, 2, \ldots, M$, using (7) and (10) we obtain

$$A(\mathcal{K}^I_{p,d}) \leq 4p^2\theta^2 d^2 \left(\log e\right)^2 = O(p^2 d). \quad (11)$$

where we use the fact that $\theta = O(\frac{1}{\sqrt{d}})$. Using (4), (6) and (11) we see that the second term of the RHS of (5) goes to 0 as $p \to \infty$ and hence probability of error goes to 1. □

Proof (Lemma 5). The upper bound on $B(s, \mathcal{G}'_{p,d})$ is obtained using arguments similar to those presented in the proof of Lemma 1. For lower bound on $\log M$, note that there are $N(p) = \frac{p!}{2^{p/2}(p/2)!}$ possible perfect matchings on a set of p nodes. Therefore, a multigraph composed of d perfect matchings can be formed in $(N(p))^d$ ways. Note that multiple copies of the same multigraph may be generated during this construction. Using Lemma 3, atleast $\left(1 - \frac{c}{p^\tau}\right)(N(p))^d$ of these multigraphs have (weighted) adjacency matrix A satisfying $\rho(A) < 3\sqrt{d}$, for some constant $c > 0$. Note that any given multigraph generated by d perfect matchings is a d-regular graph and has $\frac{pd}{2}$ edges. In general, each of the edges can come from any of the d perfect matchings. Therefore, a single multigraph can potentially be generated by atmost $d^{\frac{pd}{2}}$ sets of d perfect matchings. Also, we desire that the multigraphs have different underlying undirected graph structures in $\mathcal{G}_{p,d}$. The fact that there can be atmost d edges between two nodes of the multigraph and there are $\frac{pd}{2}$ edges overall, gives the lower bound $M = |\mathcal{H}'| = |\mathcal{K}^G_{p,d}| = |\mathcal{G}'_{p,d}| \geq \left(1 - \frac{c}{p^\tau}\right)\frac{1}{d^{pd}}(N(p))^d$. For $p > (2c)^{\frac{1}{\tau}}$,

$\left(1-\frac{c}{p^\tau}\right) \geq 1/2$. Taking log on both sides and simplifying gives

$$\log M \geq \log\left[\frac{1}{d^{pd}}\left(1-\frac{c}{p^\tau}\right)\left(\frac{p!}{2^{p/2}(p/2)!}\right)^d\right] \geq \frac{pd}{2}\log\frac{p}{4d^2} - 1,$$

for $p > (2c)^{\frac{1}{\tau}}$ and using the bound $p! \geq \left(\frac{p}{2}\right)^{\frac{p}{2}}\left(\frac{p}{2}\right)!$. □

Proof (Lemma 6). By definition we have $I(K_i; X^{(1)}) = h(X^{(1)}) - h(X^{(1)}|K = K_i)$. The differential entropy of $X^{(1)}$ is upper bounded by the differential entropy of a Gaussian random vector with the same covariance matrix. This gives $h(X^{(1)}) \leq \frac{1}{2}\log(2\pi e)^p|\bar{\Sigma}|$, where $\bar{\Sigma} = \frac{1}{M}\sum_{i=1}^M \Sigma_i$ and $\Sigma_i = \Theta_i^{-1}$ is the covariance matrix associated with K_i. Also, $h(X^{(1)}|K = K_i) = \frac{1}{2}\log(2\pi e)^p|\Sigma_i|$. This gives

$$I(K_i; X^{(1)}) \leq \frac{1}{2}(\log|\bar{\Sigma}| - \log|\Sigma_i|) = \frac{1}{2}(\log|\bar{\Sigma}| + \log|\Theta_i|)$$

By construction, the diagonal entries of every inverse covariance matrix Θ_i are same and equal to $4\sqrt{d}\mu + \delta$. So by Hadamard's Inequality, $|\Theta_i| \leq (4\sqrt{d}\mu + \delta)^p$. Also, as stated in the proof of Lemma 4, the minimum eigenvalue of Θ_i is atleast δ. This means that the maximum eigenvalue of Σ_i is atmost $\frac{1}{\delta}$ or $||\Sigma_i||_2 \leq \frac{1}{\delta}$. Hence, the maximum eigenvalue of $\bar{\Sigma}$ does not exceed $\frac{1}{\delta}$ as $||\bar{\Sigma}||_2 \leq \frac{1}{M}\sum_{i=1}^M ||\Sigma_i||_2$. This gives $|\bar{\Sigma}| \leq ||\bar{\Sigma}||_2^p \leq \frac{1}{\delta^p}$. Therefore, we obtain

$$I(K_i; X^{(1)}) \leq \frac{p}{2}\log\left(1 + \frac{4\sqrt{d}\mu}{\delta}\right) = \frac{p}{2}\log\left(1 + \frac{4\sqrt{d}}{\lambda^{-1} - 4\sqrt{d}}\right)$$

where we substitute $\mu = \frac{\delta}{\lambda^{-1} - 4\sqrt{d}}$. □

Proof (Theorem 3). Following the lines of the proof of Theorem 2, we define the bounds R and C satisfying (1) and (2), as per Lemmas 5 and 6:

$$R := \frac{pd}{2}\log\frac{p}{4d^2} - \log\left[s\binom{p^2/2}{s}\right] - 1, \tag{12}$$

$$C := \frac{p}{2}\log\left(1 + \frac{4\sqrt{d}}{\lambda^{-1} - 4\sqrt{d}}\right). \tag{13}$$

After some algebraic manipulations, we obtain the following lower bound on $\frac{R}{C}$:

$$\frac{R}{C} \geq \frac{\left(d - \frac{4s}{p}\right)\log p - 2d\log 2d + \frac{2s}{p}\log\frac{2s}{e} - \frac{2\log s}{p} - \frac{2}{p}}{\log\left(1 + \frac{4\sqrt{d}}{\lambda^{-1} - 4\sqrt{d}}\right)}.$$

Under the hypothesis of the theorem, we have $n < \frac{R}{2C}$. Theorem 1 shows that $P_e^{(n)}$ satisfies

$$P_e^{(n)} \geq 1 - \frac{8A(\mathcal{K}_{p,d}^G)}{RC} - 2^{-\frac{R}{4}}. \tag{14}$$

Even in this case, the last term in the RHS of (14) goes to 0 as $p \to \infty$, as $d = o(p^\alpha)$ for some $\alpha < \frac{1}{2}$ and $s < \frac{(1-2\alpha)pd}{8}$ implies

$$R = \Theta(pd \log p). \tag{15}$$

To show that the second last term in RHS of (14) goes to 0 as $p \to \infty$, we give a bound for $A(\mathcal{K}_{p,d}^G)$. For this, we first derive a deterministic bound on $\log \frac{p\left(X^{(1)}|K=K_i\right)}{p(X^{(1)})}$. We define $D_{max} = \max_i |\Theta_i|$, $D_{min} = \min_i |\Theta_i|$, λ_{max} to be the maximum among the eigenvalues of Θ_i, $i = 1, 2, \ldots, M$, and $\bar{\Theta} = \frac{1}{M}\sum_{i=1}^M \Theta_i$. Then given $K_i \in \mathcal{K}_{p,d}^G$ and $x \in \mathbb{R}^p$, we have the following upper bound:

$$\begin{aligned}
\frac{p(x|K=K_i)}{p(x)} &= \frac{\sqrt{|\Theta_i|}\exp(-\frac{1}{2}x^T\Theta_i x)}{\frac{1}{M}\sum_{j=1}^M \sqrt{|\Theta_j|}\exp(-\frac{1}{2}x^T\Theta_j x)} \\
&\leq \frac{\sqrt{D_{max}}}{\sqrt{D_{min}}} \frac{1}{\frac{1}{M}\sum_{j=1}^M \exp(-\frac{1}{2}x^T\Theta_j x)} \\
&\leq \frac{\sqrt{D_{max}}}{\sqrt{D_{min}}} \exp\left(\frac{1}{2}x^T\bar{\Theta}x\right) \\
&\leq \frac{\sqrt{D_{max}}}{\sqrt{D_{min}}} \exp\left(\frac{\lambda_{max}}{2}x^T x\right).
\end{aligned}$$

This gives

$$\log \frac{p(x|K=K_i)}{p(x)} \leq \frac{1}{2}\log \frac{D_{max}}{D_{min}} + \frac{\lambda_{max}}{2}x^T x. \tag{16}$$

Similarly, we have the following lower bound:

$$\begin{aligned}
\frac{p(x|K=K_i)}{p(x)} &= \frac{\sqrt{|\Theta_i|}\exp(-\frac{1}{2}x^T\Theta_i x)}{\frac{1}{M}\sum_{j=1}^M \sqrt{|\Theta_j|}\exp(-\frac{1}{2}x^T\Theta_j x)} \\
&\geq \frac{\sqrt{D_{min}}}{\sqrt{D_{max}}} \exp\left(-\frac{1}{2}x^T\Theta_i x\right) \\
&\geq \frac{\sqrt{D_{min}}}{\sqrt{D_{max}}} \exp\left(-\frac{\lambda_{max}}{2}x^T x\right).
\end{aligned}$$

This gives

$$\log \frac{p(x|K=K_i)}{p(x)} \geq -\frac{1}{2}\log \frac{D_{max}}{D_{min}} - \frac{\lambda_{max}}{2}x^T x. \tag{17}$$

Inequalities (16) and (17) together give

$$\left|\log \frac{p(X|K=K_i)}{p(X)}\right| \leq \frac{1}{2}\log\frac{D_{max}}{D_{min}} + \frac{\lambda_{max}}{2}X^TX.$$

Therefore, we have

$$\begin{aligned}\mathrm{var}\left(\log \frac{p(X|K=K_i)}{p(X)} \;\middle|\; K=K_i\right) &\leq E\left[\left(\frac{1}{2}\log\frac{D_{max}}{D_{min}} + \frac{\lambda_{max}}{2}X^TX\right)^2 \;\middle|\; K=K_i\right] \\ &\leq \frac{1}{2}\left(\log\frac{D_{max}}{D_{min}}\right)^2 + \frac{\lambda_{max}^2}{2}E[(X^TX)^2|K=K_i].\end{aligned}$$

For the given ensemble, we have $D_{max} \leq (4\sqrt{d}\mu+\delta)^p$, $D_{min} \geq \delta^p$, $\lambda_{max} = (d+4\sqrt{d})\mu+\delta \leq 5d\mu+\delta$, where $\mu = \delta/(\lambda^{-1}-4\sqrt{d})$. Using these bounds, we get

$$\begin{aligned}\mathrm{var}\left(\log \frac{p(X|K=K_i)}{p(X)} \;\middle|\; K=K_i\right) \leq \frac{p^2}{2}&\left(\log\left(1+\frac{4\sqrt{d}}{\lambda^{-1}-4\sqrt{d}}\right)\right)^2 \\ &+ \frac{\delta^2}{2}\left(1+\frac{5d}{\lambda^{-1}-4\sqrt{d}}\right)^2 E[(X^TX)^2|K=K_i].\end{aligned}$$

We also have

$$E[(X^TX)^2|K=K_i] = (Tr(\Theta_i^{-1}))^2 + Tr(2\Theta_i^{-2}) \leq \frac{p^2+2p}{\delta^2} \leq \frac{2p^2}{\delta^2}.$$

Then we get

$$\begin{aligned}\mathrm{var}\left(\log \frac{p(X|K=K_i)}{p(X)} \;\middle|\; K=K_i\right) &\leq \frac{p^2}{2}\left[\left(\log\left(1+\frac{4\sqrt{d}}{\lambda^{-1}-4\sqrt{d}}\right)\right)^2 + 2\left(1+\frac{5d}{\lambda^{-1}-4\sqrt{d}}\right)^2\right] \\ &\leq \frac{3p^2}{2}\left(1+\frac{5d}{\lambda^{-1}-4\sqrt{d}}\right)^2.\end{aligned}$$

Thus, we get the result:

$$A(\mathcal{K}_{p,d}^G) \leq \frac{3p^2}{2}\left(1+\frac{5d}{\lambda^{-1}-4\sqrt{d}}\right)^2. \tag{18}$$

For $\lambda = O\left(1/\sqrt{d}\right)$, we obtain $A = O(p^2d)$. Using (13), (15) and (18), we see that the second term of the RHS of (14) goes to 0 as $p \to \infty$ and hence probability of error goes to 1. □