Open access • Proceedings Article • DOI:10.1109/ISIT.2012.6284018

# Learning Markov graphs up to edit distance — Source link

Abhik Kumar Das, Praneeth Netrapalli, Sujay Sanghavi, Sriram Vishwanath

**Institutions:** University of Texas at Austin

**Published on:** 01 Jul 2012 - International Symposium on Information Theory

**Topics:** Edit distance, Markov property, Variable-order Markov model, Hammersley–Clifford theorem and Markov chain

# Learning Markov Graphs Up To Edit Distance

Abhik Kumar Das, Praneeth Netrapalli, Sujay Sanghavi and Sriram Vishwanath
Department of ECE, The University of Texas at Austin, USA
e-mail:{akdas,praneethn}@utexas.edu, sanghavi@mail.utexas.edu, sriram@ece.utexas.edu

***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. We first prove a general result for any probability distribution, and then specialize it for Ising and Gaussian models. 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.**

***Index Terms*—Markov networks, graphical models, strong converse, Ising model, Gaussian Markov model**

## I. Introduction

Markov networks, also known as (undirected) graphical models, describe the interdependencies (or the lack 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 probability distribution is a well-studied one and is referred to as the problem of graphical model selection or Markov graph learning. There is diverse literature on various aspects of learning Markov graphs, ranging from statistical physics to computational learning theory. It is only relatively recently that the information theoretic limits for this learning problem are being better understood.

An understanding of the information theoretic limits of high-dimensional learning problems in general, and the problem of learning Markov graphs 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 the problem of learning Markov graphs is that if the number of observed samples available to a learning algorithm falls below a certain threshold, the probability of error in learning the structure of a Markov network is bounded away from zero. Therefore, alternate information-theoretic tools are required for stronger bounds on sample complexity. This motivates 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 that say that unless the number of available samples exceeds some threshold, the probability of error in learning the structure of a Markov network goes to one as the problem size increases. Such information-theoretic limits are important as they provide an understanding of the settings where recovery is impossible, regardless of the algorithm or 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 results for two well-known families of Markov networks – Ising models and Gaussian Markov networks. The distortion metric we choose is edit distance between graphs, that we define in the next section.

### A. 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 Erdös-Rényi graph based Ising models and degree bounded Gaussian Markov networks respectively. The performance of graphical model learning algorithms, that output a list of graph structures instead of a single one, is examined in [15] for Gaussian and Ising models.

### B. Summary of Results

We provide a comparison of our results against the existing ones in literature in Table I, for ensemble of Markov networks based on graphs with $p$ nodes and degree bounded by $d$. Our results are highlighted in bold face. Note that $s$ denotes the maximum allowed edit distance between the original and recovered graphs. All existing results in literature are for the case of exact graph structure recovery i.e., $s = 0$.

TABLE I
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]<br>$\boldsymbol{\Omega\left(\left(d - \frac{8s}{p}\right) \log p\right)}$ | $\Omega(\sqrt{d}e^{\sqrt{d}} \log p)$ [9]<br>$\boldsymbol{\Omega\left(\left(d - \frac{8s}{p}\right) \log p\right)}$ |
| Gaussian | $\Omega\left(d^2 \log p\right)$ [11]<br>$\boldsymbol{\Omega\left(\sqrt{d}\left(d - \frac{4s}{p}\right) \log p\right)}$ | $\Omega\left(d \log p\right)$ [11]<br>$\boldsymbol{\Omega\left(\left(d - \frac{4s}{p}\right) \log p\right)}$ |

It is known that edge weights play an important role in determining the complexity of learning graphical models [16]. However this dependency is complex in general. For instance, very low edge weights or very large edge weights tend to increase the number of samples needed to learn the graphical model. The presence of low edge weights causes difficulty in determining the existence of edges, while the presence of high edge weights results in long range correlations in graphical models. Existing results show significant difference in lower bounds for edge weight scaling as $\Theta(\frac{1}{d})$ as opposed to $\Theta(\frac{1}{\sqrt{d}})$, see Table I. However, in this paper, we are able to show different lower bounds for the Gaussian case but not for the Ising model. A detailed comparison and discussion of our results with the existing ones is done in Section VI.

The rest of this paper is organized as follows. We discuss some preliminaries and introduce the system model in Section II. 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 III, IV and V respectively. We finally conclude the paper with Section VI. Due to lack of space, complete proofs of all the results are not presented in this paper. They can be found in [17].

## II. PRELIMINARIES

We consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes and $E \subseteq V \times V$ is the set of edges. A Markov network is obtained by associating a random variable $X_i$ to node $i$, that takes values from an alphabet set $\mathcal{A}$, and specifying a joint probability distribution $p(\cdot)$ 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 $x_A, x_B, x_C$ denote the restrictions of $(x_1, \ldots, x_p) \in \mathcal{A}^p$ to indices in $A, B, C$ respectively. Here, $p(\cdot)$ denotes p.m.f. for the discrete case ($\mathcal{A}$ is a finite set) and p.d.f. for the continuous case ($\mathcal{A} = \mathbb{R}$ and continuous distribution). Next, we present examples of discrete and continuous Markov networks.

**Ising Model:** This family of discrete probability distributions is widely studied and used in statistical physics [18], computer vision [19], game theory [20] and many other topics. In this paper, we restrict ourselves to a special case of Ising model – the *zero field* binary Ising model. Here, the alphabet is chosen as $\mathcal{A} = \{-1, 1\}$. Given an undirected graph $G$ on $p$ nodes and weight $\theta_{ij} \in \mathbb{R}$ assigned to edge $(i, j) \in E$, the probability of vector $x = (x_1, x_2, \ldots, x_p) \in \mathcal{A}^p$ is given by

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

where vector $z = (z_1, z_2, \ldots, z_p)$ ranges over $\mathcal{A}^p$.

**Gaussian Markov Model:** This is one of the most well known families of continuous Markov networks. Here, vector $X$ possesses a multivariate Gaussian distribution over reals. Without loss of generality, it can be assumed that $X$ has the zero vector as its mean. Given an undirected graph $G$ on $p$ nodes and a $p \times p$ positive definite matrix $\Theta \in \mathbb{R}^{p \times p}$ such that $\Theta(i, j) \neq 0$ iff $(i, j) \in E$, the p.d.f. of vector $X$ is given by

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

where $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and $\Theta$ is the inverse covariance matrix. $\Theta$ 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 the minimum magnitude of partial correlation coefficient, plays an important role in determining the sample complexity of the Markov graph 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 $\Theta$.

In this paper we restrict our attention to the ensemble of degree bounded graphs in light of the fact that extensive work on learning Markov networks focus on these graphs. We denote the set of 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*, $\Delta(G, H)$, between them as the minimum number of edge deletions/insertions required to convert $G$ to $H$. Thus, $\Delta(G, H)$ is the cardinality of the symmetric difference between edge sets of $G$ and $H$.

### A. 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 $K_i$ has $G_i$ as its underlying graph and the random variables in $X = (X_1, \ldots, X_p)$ 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 referred to as the problem of graphical model selection or Markov graph learning. A learning algorithm is any function $\phi : \mathcal{A}^{np} \to \mathcal{U}_p$ that maps the observed samples to an estimated graph $\hat{G} = \phi(X^n) \in \mathcal{U}_p$. Given a pre-specified $s$, we define the error event for the learning algorithm as $\{\Delta(G, \phi(X^n)) \geq s\}$, i.e., error is declared if the edit distance between the actual graph and the reconstructed version is greater than or equal to $s$. Then the probability of error of the learning algorithm is given by the following expression:

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

In this paper, we derive lower bounds on 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 $P_e^{(n)}$ from below in terms of $n$ and the ensemble parameters.

## III. 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 II-A. We have an ensemble of Markov networks $\mathcal{K}$ and the corresponding ensemble of undirected graphs $\mathcal{G}$ on $p$ nodes. We choose $K \in \mathcal{K}$ uniformly at random and obtain $n$ i.i.d. sample vectors from its joint distribution. Our aim is to analyze the performance of an arbitrary learning algorithm $\phi : \mathcal{A}^{np} \to \mathcal{U}_p$. For this, we define the following quantities for $G \in \mathcal{G}$:

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

$B(s, \mathcal{G})$ denotes the maximum number of graphs 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:

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

$H(\cdot)$ and $h(\cdot)$ represent the entropy and differential entropy functions respectively. For given $\mathcal{K}, \mathcal{G}$, 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 the following theorem establishes a necessary condition on the number of samples $n$ for consistent recovery of the structure of Markov networks 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:*

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

*Here, error refers to the event that the original graph and the recovered graph are at an edit distance of more than s, and*

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

*where $p(\cdot)$ stands for p.m.f. in the discrete case and p.d.f. in the continuous case, and dependent on ensemble $\mathcal{K}$.*

If we can find a good upper bound for $A\left(\mathcal{K}\right)$ for given ensemble $\mathcal{K}$, then 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.

## IV. Lower Bounds for Ising Models

Our main result in this section is the following theorem that characterizes a lower bound on the number of samples required for consistent recovery of the Markov graph structure for an ensemble of Ising models whose construction is described below. For each graph in $\mathcal{G}_{p,d}$, we consider the corresponding zero field binary Ising model with all edge weights equal to $\theta$, where $\theta \in (0, \frac{1}{\sqrt{d}})$. We refer to this ensemble as $\mathcal{K}_{p,d}^I$. Note that there is a bijective mapping between $\mathcal{G}_{p,d}$ and $\mathcal{K}_{p,d}^I$.

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

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

*then for any arbitrary graphical model learning algorithm, its probability of error satisfies $P_e^{(n)} \to 1$ 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 and [17]. We list the graphs in $\mathcal{G}_{p,d}$ as $\mathcal{G}_{p,d} = \{G_1, \ldots, G_M\}$ 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 these ensembles.

**Lemma 1.** *For $\mathcal{G}_{p,d}$ with $d \leq \frac{p-1}{2}$, $s \leq \frac{p(p-1)}{4}$, we have*

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

**Lemma 2.** *Suppose $K$ is chosen uniformly at random from $\mathcal{K}_{p,d}^I = \{K_1, \ldots, K_M\}$. Then we have the following bound:*

$$\max_{1 \leq i \leq M} I(K_i; X^{(1)}) \leq p.$$

## V. Lower Bounds for Gaussian Markov Models

Our main result in this section is a lower bound on the number of samples required for consistent recovery of the Markov graph structure for an ensemble of Gaussian Markov networks $\mathcal{K}^G_{p,d}$, whose construction is described below.

Without any 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 edges in the matchings. We refer to the set of all such multigraphs on $p$ nodes, constructed in this fashion, as $\mathcal{H}$. The uniform distribution over the choices of perfect matchings also defines a probability distribution over $\mathcal{H}$. We have the following lemma for this distribution [21]:

**Lemma 3.** *Consider a multigraph $H$ on $p$ nodes, formed from the union of $d$ random perfect matchings on the nodes that are 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 constant and $\tau = \lceil \frac{\sqrt{d-1}+1}{2} \rceil - 1$.*

Next, we eliminate those multigraphs from $\mathcal{H}$ whose (weighted) adjacency matrices $A$ satisfy $\rho(A) \geq 3\sqrt{d}$ and get a reduced subset $\mathcal{H}'$. By Lemma 3, $\mathcal{H}\backslash\mathcal{H}'$ forms a small fraction of $\mathcal{H}$. We fix constants $\lambda \in (0, \frac{1}{4\sqrt{d}})$, $\delta > 0$ 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.** *Every $\Theta \in \mathcal{T}$ is symmetric and positive definite.*

*Proof:* By construction, given any $\Theta \in \mathcal{T}$, we have $\Theta = (4\sqrt{d}\mu + \delta)I_p + \mu A$, 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 $\Theta$ is at least $4\sqrt{d}\mu + \delta - \rho(A)\mu > \sqrt{d}\mu + \delta > 0$. This along with the symmetry of $\Theta$ ensure that all the eigenvalues of $\Theta$ are positive. Hence, $\Theta$ is also a positive definite matrix. ■

Note that the choice of $\mu$ ensures $\lambda^*(\Theta) = \lambda$ for every $\Theta \in \mathcal{T}$. Lemma 4 suggests that the matrices of $\mathcal{T}$ can potentially 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 graphical 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 for some $\alpha < \frac{1}{2}$, $d = o(p^\alpha)$, $s < (1-2\alpha)\frac{pd}{8}$ and the number of samples, obtained from distribution of $K$, satisfies*

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

*then for any arbitrary graphical model learning algorithm, its probability of error satisfies $P_e^{(n)} \to 1$ 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. The complete proof of Theorem 3 can be found in the appendix and [17]. We list the graphs in $\mathcal{G}'_{p,d}$ as $\mathcal{G}_{p,d} = \{G_1, \ldots, G_M\}$ and the corresponding Gaussian Markov networks in $\mathcal{K}^G_{p,d}$ as $\{K_1, \ldots, K_M\}$. Then the following two lemmas hold.

**Lemma 5.** *For $\mathcal{G}'_{p,d}$ with large enough $p$, $s \leq \frac{p(p-1)}{4}$, we have*

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

**Lemma 6.** *Suppose $K$ is chosen uniformly at random from $K^G_{p,d} = \{K_1, \ldots, K_M\}$. Then we have the following bound:*

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

## VI. Conclusion & Discussion

**Remarks about Theorems 2 & 3:** Specializing our result to the case of exact recovery, i.e., $s = 0$ yields weaker lower bounds than the existing results. However, for the case of Gaussian Markov models [11] with edge weights $\Theta(\frac{1}{\sqrt{d}})$, our result matches the existing result and for both the cases of Ising [9] and Gaussian [11] models with edge weights $\Theta(\frac{1}{d})$, 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 the probability of error goes to 1 as $p \to \infty$.

In this paper, we develop a rate-distortion framework for the problem of learning Markov graphs, 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 fewer samples. Our results show that, for both Ising and Gaussian models on $p$ variables with maximum degree $d$, at least $\Omega((d - \frac{s}{p})\log p)$ samples are required for any algorithm to recover the graph within edit distance $s$. As future work, we hope to derive stronger bounds on probability of error in learning, and derive results for Markov networks based on other graph ensembles, like random graphs.

## 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 Info. Theory*. Wiley Interscience, 2006.

[8] R. Gallager, *Info. 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] A. Das, P. Netrapalli, S. Sanghavi, and S. Vishwanath, "Learning Markov graphs up to edit distance," http://uts.cc.utexas.edu/∼akdas/mglearning_techreport12.pdf, 2012.

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

[19] 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.

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

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

## Appendix

### A. Proof of Theorem 2

We make use of Lemmas 1 and 2 to construct lower bound $R$ and upper bound $C$, as defined 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 $\binom{a}{b} \leq \left(\frac{a\cdot e}{b}\right)^b$, we obtain the following lower bound:

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

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

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

We need to show that the last two terms of the RHS of (5) go to 0 as $p \to \infty$. Since $d = o(p^\alpha)$ 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 term of the RHS of (5), we use the following bound whose proof is given in [17]:

$$A(\mathcal{K}^I_{p,d}) \leq 4p^2\theta^2 d^2\left(\log e\right)^2 = O(p^2 d). \tag{7}$$

Here, we use the fact that $\theta = O(\frac{1}{\sqrt{d}})$. Using (4), (6) and (7) 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.

### B. Proof of Theorem 3

Following the approach of proof of Theorem 2, we define lower bound $R$ and upper bound $C$, given by (1) and (2), as:

$$R := \frac{pd}{2}\log\frac{p}{4d^2} - \log\left(s\binom{p^2/2}{s}\right) - 1, \tag{8}$$

$$C := \frac{p}{2}\log\left(1 + \frac{4\sqrt{d}}{\lambda^{-1} - 4\sqrt{d}}\right). \tag{9}$$

Lemmas 5 and 6 show that $R$ and $C$ satisfy (1) and (2). After some simplifications, we obtain the following lower bound:

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

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

$$P_e^{(n)} \geq 1 - \frac{8A(\mathcal{K}^G_{p,d})}{RC} - 2^{-\frac{R}{4}}. \tag{10}$$

We need to show that the last two terms of the RHS of (10) go to 0 as $p \to \infty$. To constrain the second term we use the following upper bound whose proof can be found in [17]:

$$A(\mathcal{K}^G_{p,d}) \leq \frac{3p^2}{2}\left(1 + \frac{5d}{\lambda^{-1} - 4\sqrt{d}}\right)^2. \tag{11}$$

For $\lambda = O(\frac{1}{\sqrt{d}})$, we obtain $A(\mathcal{K}^G_{p,d}) = O(p^2 d)$. Also note that since $d = o(p^\alpha)$ for some $\alpha < \frac{1}{2}$ and $s < \frac{(1-2\alpha)pd}{8}$, we have

$$R = \Theta(pd\log p). \tag{12}$$

Using (9), (11) and (12), we see that the last two terms of the RHS of (10) goes to 0 as $p \to \infty$, hence $P_e^{(n)} \to 1$.