

Sample Complexity of Determining Structures of Graphical Models

Narayana Santhanam¹ Martin J. Wainwright²
^{1,2}EECS and ²Statistics Departments, UC Berkeley, Berkeley, CA 94720

Abstract—We consider discovering the graph structure of a pairwise Markov random field (MRF) on p binary random variables using n samples from the underlying MRF distribution. We analyze the information-theoretic limitations of this problem under high-dimensional scaling, when the number of connections of each variable in the underlying MRF is bounded by d . We derive both necessary and sufficient conditions on the scaling of the triplet (n, p, d) for asymptotically reliable recovery of the graph structure.

I. INTRODUCTION

The problem of *graphical selection* is to correctly estimate the graph structure of a Markov random field given samples from the underlying distribution. This problem is central to statistics and machine learning, with consequences for a variety of application domains, e.g., image analysis [2], social networks, and computational biology [6].

In general, a graphical model is a structured representation of a joint distribution of potentially dependent random variables. In the models considered in this paper, we use the vertices of a graph $G = (V, E)$ to represent our (binary) random variables. Let $|V| = p$. Corresponding to the graph, we associate a joint probability distribution on the p random variables with the edges E determining the exact joint distribution. Perhaps a big strength of the graphical model representation is the visualization possible about which models are close to each other. One would expect distributions with similar graphs to be close to each other and indeed, as an auxiliary result, we formalize this notion in this paper.

The distributions we consider are *Markov random fields* with respect to their graphs. Namely, each distribution satisfies a spatial Markov-type independence relationship—any random variable X , conditioned on those adjacent to it in the graph associated with the distribution, is independent of all other variables.

Consider a set \mathcal{G} of such models. A sample from a distribution in \mathcal{G} is a p -dimensional binary vector, corresponding to the realization of the p random variables. Given several samples from a distribution, formally, the graph selection problem requires us to choose the structure of the model in \mathcal{G} that best fits the data.

If the underlying graph is known to be tree-structured, then the problem can be reformulated as a maximum-

weight spanning tree problem [5], and solved in polynomial time. On the other hand, for fully general graphs with cycles, the problem is known to be computationally hard [4]. Nonetheless, a variety of practical methods have been proposed, including constraint-based approaches [14], heuristic search, and ℓ_1 -based relaxations [9], [16], [17]. In particular, it can be proven that some of the ℓ_1 -based approaches are consistent for model selection under particular scalings of the graph size, degrees, and number of samples [9], [16], and for the case where each degree is bounded by d , the sample complexity has been independently obtained in [3].

Of complementary interest—and the focus of the paper—are the information-theoretic limitations of the graphical selection problems for pairwise binary Markov random fields, also known as *Ising models* [1]. The Ising model (1) has its origins in statistical physics [1], where it is used to model physical phenomena such as crystal structure and magnetism. Its relative simplicity and the ability to capture practically useful dependencies makes it a candidate for image processing [2], [7], gene network analysis, analysis of social networks, and recently in coding theory starting from [13] (see e.g. [10] for details and a list of references on the topic) and communications, e.g. [8].

At the same time, the dimensionality of the data sets—the number of random variables in question—has increased significantly for many practical problems. For example, gene networks attempt to model interactions between the expression levels of thousands of genes, usually using only tens of samples. The natural question, one that will be addressed here, is whether it is even possible to obtain anything useful from such limited data, and if so, what?

We are given n i.i.d. samples—namely, n p -dimensional vectors—from a fixed but unknown model in \mathcal{G} . As mentioned before, we consider inferring E of the graph alone—namely, given samples from an unknown model in \mathcal{G} , we have to infer the structure (edge set E) of the unknown model. This setting is useful for problems in gene regulatory network inference, for example, where there is a severe shortage of data to do a model selection problem.

How must n scale with parameters corresponding to

the graphs of the \mathcal{G} so that there exists a method that recovers the correct probability model? Conversely, for what scalings does any method fail to recover the underlying structure correctly? Note that from this perspective, the graphical selection problem is a channel coding problem, in which the messages are the graphs in our family, and each use of the channel corresponds to taking an i.i.d. sample.

We analyze the information-theoretic limitations of this problem under high-dimensional scaling, when the connections of each variable in the underlying MRF is bounded by d . We derive both necessary and sufficient conditions on the scaling of the triplet (n, p, d) for asymptotically reliable recovery of the graph structure.

In Section II, we set up the problem of determining the sample complexity of graphical selection formally. Section III contains the formal statements of our sample complexity results. Section VI contains our results regarding the divergence between Ising models, showing that, roughly speaking, distributions represented by graphs that are similar are indeed closer. Section V outlines the core results that we use for the proofs of results in Section III, and Sections VII and VIII outline the proofs.

II. PROBLEM STATEMENT

We begin with a precise statement of the problem.

Given an undirected graph $G = (V, E)$, let us associate with each vertex $i \in V$ a binary random variable $X_i \in \{-1, +1\}$. We then consider a distribution over the random vector $X = (X_1, \dots, X_p)$ with probability mass function

$$\mathbb{P}_\lambda(x) = \frac{1}{Z(\lambda)} \exp \left\{ \sum_{(s,t) \in E} \lambda_{st} x_s x_t \right\}, \quad (1)$$

where $Z(\lambda)$ is the normalization constant.

We will refer to the $\binom{p}{2}$ parameters, $\lambda_{\{st\}}$, $s, t \in V$, as the *edge parameters*.

At the outset, the sample complexity of inferring an edge set depends in on some form of an upper bound on the edge parameters. As the following example indicates, if the edge parameters are unboundedly large, it may be impossible to distinguish distinct models (potentially with distinct edge sets), even given infinite amounts of data.

Example 1. Consider the set of all graphical models on $p = 3$ variables, and every edge parameter $\in \{0, \lambda\}$. Note that there are a total of three such graphs corresponding to graphs on 3 vertices with exactly 2 edges. If $\lambda = \infty$, it is easy to see that all these models reduce to the *guilt by association* distribution—i.e., the two configurations $\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$ and $\begin{bmatrix} -1 & -1 & -1 \end{bmatrix}$ have

probability $\frac{1}{2}$ each. There is hence no way to distinguish between the 3 models in the case $\lambda = \infty$. \square

To take into account the above observation, our results for a class \mathcal{G} depend on B —the smallest real number that satisfies the following for all $\Lambda \in \mathcal{G}$. If Λ is associated with the graph (V, E) , with parameters λ_{st} on $(s, t) \in E$, then for all $s \in V$,

$$\sum_{(s,t) \in E} |\lambda_{st}| \leq B. \quad (2)$$

No matter what the class \mathcal{G} is, selection of a graphical model is not necessarily the same as simply estimating pairs of vertices (s, t) for which $\text{cov}\{X_s, X_t\} \neq 0$; indeed, given the distribution (1), it can be seen that X_s and X_t could be correlated even when there is no direct edge connecting them.

On the other hand, if each edge parameter that is non-zero is finite, given infinite amounts of data, two distinct models $\Lambda^{(i)}$ and $\Lambda^{(j)}$ can always be told apart by looking at the vector of all covariances $\{EX_s X_t\}_{\substack{(s,t) \in V^2 \\ s \neq t}}$.

We observe that a new elementary proof of the above statement can be constructed along the lines of the arguments presented here (see [11] for full proofs). Moreover, we note that a more general result on exponential families exists along these lines, see [15].

With finite data covariances can only be approximated. The question therefore is, how fine should these approximations be? This question determines the number of samples that are necessary for the graph selection problem, and a constructive argument determines the number of samples that are sufficient.

In [12], we considered fitting bounded degree models. Let θ be a $\binom{p}{2}$ length vector of real numbers. We index the components of the vector by a pair of numbers $\{s, t\}$, $s, t \in V$, and write the components as θ_{st} . Furthermore, assume that for some real number $\lambda > 0$, $|\theta| \geq \lambda \mathbf{1}$, where the vector inequality is taken to mean a component-wise one. Then the set $\mathcal{G}_{d,p,\theta}$ contains all models on vertices V such that

- (a) each vertex v in V has degree at most d , namely for the edge set E associated with the model, $|\{e \in E : v \in e\}| \leq d$, for some $d \geq 1$;
- (b) if $(s, t) \in E$, the edge parameter is θ_{st} , else 0.

We build on the above case to let the edge parameters be arbitrary, as long as they are bounded. Specifically,

$$\mathcal{G}_{d,p} = \cup \mathcal{G}_{d,p,\theta},$$

where the union is taken over all θ such that $|\theta| \geq \lambda \mathbf{1}$ and each class in the union satisfies Equation (2). Note that corresponding to each edge set, there is now a (infinite) set of models possible. The task now is to determine the edge set of the underlying model, hence it is sufficient to

distinguish between the subsets of $\mathcal{G}_{d,p}$ that correspond to different edge sets, rather than different models.

One observation regarding the models described above: in a model belonging to any of the classes above, for all vertices $s \in V$, $EX_s = 0$.

III. RESULTS

We begin by stating the results obtained for the sample complexity of inferring the structure of graphical models. The following theorem addresses the sufficiency part.

Theorem 1. *Let $0 < \delta \leq 1$ and let*

$$\lambda = \min_{(s,t) \in V^2} \lambda_{st}.$$

If

$$n > \frac{(3e^{2B} + 1)^2}{\sinh^4\left(\frac{\lambda}{2}\right)} \left(2 \log p + \log \frac{1}{\delta}\right)$$

then \exists decoder $q : X^n \rightarrow \mathcal{G}_{d,p}$ such that for all $\Lambda^{(i)}$, $1 \leq i \leq N$,

$$\mathbb{P}(q(X^n) \neq \Lambda^{(i)} | \Lambda^{(i)}) < \delta. \quad \square$$

The following theorem addresses the number of samples that are necessary for the estimation of the edge structure.

Theorem 2. *For all decoders $q : X^n \rightarrow \mathcal{G}_{d,p}$, if $B \geq 1$ and*

$$n \leq \frac{1}{2} \frac{e^{B/2} d \lambda (\log p d - 1)}{32(e^\lambda - e^{-\lambda})} + \frac{1}{4} d \log \frac{p}{d}$$

there \exists model $\Lambda^{(i)}$ such that

$$\mathbb{P}(q(X^n) \neq \Lambda^{(i)} | \Lambda^{(i)}) \geq \frac{1}{2}.$$

IV. PRELIMINARIES

To quantify how far apart and therefore, how easy it is to tell models apart, we define distance measures between models. In Section IV-B, we examine the behavior of the edge inference problem for various values of the parameter B .

A. Properties of the distance measures

We use the following distance between models $\Lambda^{(i)}$ and $\Lambda^{(j)}$

$$\mathcal{D}(\Lambda^{(i)}, \Lambda^{(j)}) \stackrel{\text{def}}{=} D\left(\frac{\Lambda^{(i)} + \Lambda^{(j)}}{2} \parallel \Lambda^{(i)}\right) + D\left(\frac{\Lambda^{(i)} + \Lambda^{(j)}}{2} \parallel \Lambda^{(j)}\right), \quad (3)$$

where the $(\Lambda^{(i)} + \Lambda^{(j)})/2$ denotes the model obtained by averaging the parameters of the two models. This bounds the Chernoff exponent in the hypothesis test between

models $\Lambda^{(i)}$ and $\Lambda^{(j)}$. In Theorem 3, we bound $\mathcal{D}(i, j)$ for any two models.

Another useful measure of distance we consider is the symmetrized KL distance between two models, $\Lambda^{(i)}$ and $\Lambda^{(j)}$,

$$\mathcal{J}(\Lambda^{(i)}, \Lambda^{(j)}) \stackrel{\text{def}}{=} D(\Lambda^{(i)} \parallel \Lambda^{(j)}) + D(\Lambda^{(j)} \parallel \Lambda^{(i)}). \quad (4)$$

Clearly $\mathcal{D}(\Lambda^{(i)}, \Lambda^{(j)}) \geq 0$ and $\mathcal{J}(\Lambda^{(i)}, \Lambda^{(j)}) \geq 0$. For the above defined distances between models $\Lambda^{(i)}$ and $\Lambda^{(j)}$, we write $\mathcal{D}(i, j)$ and $\mathcal{J}(i, j)$ where there is no ambiguity. Note that

$$\begin{aligned} \mathcal{J}(i, j) &= \mathcal{J}(\Lambda^{(i)}, \Lambda^{(j)}) \\ &= \sum_{(s,t) \in E_i \cup E_j} (\lambda_{st}^i - \lambda_{st}^j) (E_{\Lambda^{(i)}} X_s X_t - E_{\Lambda^{(j)}} X_s X_t) \end{aligned} \quad (5)$$

Corollary 1. Let $\Lambda^{(1)}$ be any Ising model. Let $\Lambda^{(2)}$ be the model obtained by increasing (decreasing) the parameter corresponding to any edge. The correlation on the edge increases (decreases) in model $\Lambda^{(2)}$ since $\mathcal{J}(1, 2) \geq 0$. \square

Lemma 1. *For all models $\Lambda^{(i)}$ and $\Lambda^{(j)}$, $\mathcal{D}(i, j) \leq \frac{1}{2} \mathcal{J}(i, j)$.*

Proof The lemma follows by noting that

$$\mathcal{D}(i, j) \leq \mathcal{J}\left(i, \frac{i+j}{2}\right) + \mathcal{J}\left(j, \frac{i+j}{2}\right),$$

and with a little bit of algebra that

$$\mathcal{J}\left(i, \frac{i+j}{2}\right) + \mathcal{J}\left(j, \frac{i+j}{2}\right) = \frac{1}{2} \mathcal{J}(i, j). \quad \square$$

B. Inference for various values of B

Let K_m be a fully connected graph on $m+1$ vertices. Let Λ_{st} be a model with its graph being K_r with the edge (s, t) removed, and every edge parameter being λ . We compute the indirect correlation on the missing edge. In the Ferromagnetic case, the FKG inequality implies that this is the maximum possible indirect correlation due to a model with any graph on these $m+1$ nodes, and edge parameters being λ .

Lemma 2. *If $B \geq 1$, for the model Λ_{st} defined above*

$$E_{\Lambda_{st}} X_s X_t \geq 1 - \frac{2(m+1)e^{3\lambda/2}}{e^{B/2} + (m+1)e^{3\lambda/2}}$$

Proof A simple computation yields

$$\begin{aligned} \frac{\mathbb{P}(X_s X_t = 1)}{\mathbb{P}(X_s X_t = -1)} &= \\ &= \frac{\sum_{l=0}^m \binom{m}{l} \exp\left(\frac{\lambda}{2} [(2l - m + 1)^2 - 4]\right)}{\sum_{l=0}^m \binom{m}{l} \exp\left(\frac{\lambda}{2} [(2l - m)^2]\right)} \end{aligned}$$

We lower bound the above ratio. To do so, we pick the largest term in the denominator. It will be shown that for $B \geq \frac{1}{2} \log m$, the largest term is the one corresponding to $i = m$ and $i = 0$, while for $1 \leq B \leq \frac{1}{2} \log m$, the largest term lies in the range $i > 3m/4$ and $i < m/4$.

Let the largest term be l^* , wlog, $\geq m/2$. It follows that

$$\begin{aligned} \frac{\mathbb{P}(X_s X_t = 1)}{\mathbb{P}(X_s X_t = -1)} &\stackrel{(a)}{\geq} \frac{\binom{m}{l^*} \exp\left(\frac{\lambda}{2}[(2l^* - m + 1)^2 - 4]\right)}{(m+1)\binom{m}{l^*} \exp\left(\frac{\lambda}{2}[(2l^* - m)^2]\right)} \\ &= \frac{\exp\left(\frac{\lambda}{2}[4l^* - 2m - 3]\right)}{m+1} \\ &\geq \frac{\exp\left(\frac{\lambda}{2}[m - 3]\right)}{m+1} \\ &= \frac{\exp\left(\frac{B}{2} - \frac{3}{2}\lambda\right)}{m+1}. \end{aligned} \quad \square$$

Consider two models in $\mathcal{G}_{d,p,\theta}$ with all their parameters being λ : (i) $\Lambda^{(1)}$ with its graph being K_d with an edge, say (s, t) , removed and (ii) $\Lambda^{(2)}$ with its graph being K_d with a different edge, say (s', t') , removed. From the FKG inequality and noting that model $\Lambda^{(2)}$ contains the edge (s, t) , observe that

$$\frac{\mathbb{P}_2(X_s X_t = 1)}{\mathbb{P}_2(X_s X_t = -1)} \leq \frac{\mathbb{P}_1(X_s X_t = 1)}{\mathbb{P}_1(X_s X_t = -1)} e^{2\lambda}.$$

Hence

$$\mathcal{J}(1, 2) \leq \frac{2(e^{2\lambda} - 1)}{\frac{\mathbb{P}_1(X_s X_t = 1)}{\mathbb{P}_1(X_s X_t = -1)}}. \quad (6)$$

Hence if $B > 1$, it follows that

$$\begin{aligned} \mathcal{J}(1, 2) &= \lambda(E_{\Lambda^{(2)}} X_s X_t - E_{\Lambda^{(1)}} X_s X_t) + \\ &\quad \lambda(E_{\Lambda^{(1)}} X_{s'} X_{t'} - E_{\Lambda^{(2)}} X_{s'} X_{t'}) \\ &\leq \frac{2Be^{5\lambda/2} \sinh(\lambda)}{e^{B/2}}. \end{aligned}$$

Similarly considering $K_{\sqrt{2k}}$ and constructing two models from it as above, we obtain two models in $\mathcal{G}_{k,p,\theta}$ separated by a distance that is exponentially small in B .

It follows that if $B \geq 1$ and any model is to be inferred with high probability, the sample complexity should grow exponentially in the parameter B .

Lemma 3. If $B \leq \frac{1}{2}$, then

$$E_{\Lambda_{st}} X_s X_t \leq 1 - \frac{2}{e^{2B} + 1}$$

Proof From the definition of B , it follows that

$$\frac{\mathbb{P}(X_s X_t = 1)}{\mathbb{P}(X_s X_t = -1)} \leq e^{2B},$$

implying the lemma.

V. TECHNIQUES

We briefly survey some of the results that form the core of the attack on Theorems 1 and 2 in the next two subsections respectively.

A. Achievable number of samples using an optimal decoder

Roughly speaking, given distinct models $\Lambda^{(i)}$ and $\Lambda^{(j)}$, there exist edge(s) (a, b) , $a, b \in V$ such that $\mathbb{E}_{\Lambda^{(i)}}[X_a X_b] \neq \mathbb{E}_{\Lambda^{(j)}}[X_a X_b]$. The sample complexity of estimating a model in $\mathcal{G}_{d,p,\theta}$ or $\mathcal{G}_{d,p}$ depends on how small the sample variance should be so that we can notice the above difference in means in a statistically significant manner.

We estimate the difference in means by using the pairwise KL divergence between models (the two-way hypothesis testing step).

B. Necessary number of samples

To estimate the necessary number of samples for decoding the edge structure, we fall back on a version of Fano's inequality. Essentially, Fano's inequality quantifies how much information is gleaned from any single sample, and therefore the number of samples we need before we have sufficient information to infer the edge structure.

1) *Fano's inequality:* We formalize the above intuition in the following lemma.

Lemma 4. For any set of w models $\Lambda^{(1)}$ through Λ_w and all decoders $q: X^n \rightarrow \{\Lambda^{(1)}, \dots, \Lambda_w\}$, if

$$n \leq \frac{w^2 \log \frac{w}{4}}{2 \sum_{1 \leq i < j \leq w} \mathcal{J}(i, j)},$$

then

$$\max_{1 \leq i \leq w} \mathbb{P}\left(q(X^n) \neq \Lambda^{(i)} | \Lambda^{(i)}\right) \geq \frac{1}{2}.$$

Proof Let Λ be a uniform random variable taking values from the set of models $\Lambda^{(1)}$ through Λ_w . Let

$$P_e = \frac{1}{w} \sum_{1 \leq i \leq w} \mathbb{P}\left(q(X^n) \neq \Lambda^{(i)} | \Lambda^{(i)}\right).$$

From Fano's inequality, $H(\Lambda | X^n) \leq 1 - P_e \log(w - 1)$, we obtain

$$\begin{aligned} &\log(w) - 1 - P_e \log(w - 1) \\ &\leq H(\Lambda) - H(\Lambda | X^n) \\ &= H(X^n) - H(X^n | \Lambda) \\ &= \frac{n}{w} \sum_{i=1}^w D(\Lambda^{(i)} || \frac{1}{w} \sum_{j=1}^w \Lambda^{(j)}) \\ &\leq \frac{n}{w^2} \sum_{1 \leq i, j \leq w} D(i || j) \\ &= \frac{n}{w^2} \sum_{1 \leq i < j \leq w} \mathcal{J}(i, j). \end{aligned}$$

The lemma follows. \square

VI. DIVERGENCE BETWEEN MODELS

We require a bound on the distance $\mathcal{D}(i, j)$ between models $\Lambda^{(i)}$ and $\Lambda^{(j)}$ in order to see how far apart the vector of edgewise correlations of the two distributions are.

It is intuitive that the distance $\mathcal{D}(i, j)$ between models should be related to the number of edges that exist only in one model. Indeed, Theorem 3 confirms this— $\mathcal{D}(i, j)$ is proportional to the *matching number* of $(E_i - E_j) \cup (E_j - E_i)$. Recall that a *matching* of a graph G is a subgraph H of G such that each vertex in H has degree 1, and that the matching number of G is the largest possible number of edges in any matching.

Theorem 3. *Let $\Lambda^{(i)}$ and $\Lambda^{(j)}$ be models with distinct edges, and let m be the matching number of the graph $(E_i - E_j) \cup (E_j - E_i)$. Let*

$$\lambda = \min_{(s,t) \in E_i \cup E_j} \lambda_{st}.$$

Then

$$\mathcal{D}(i, j) \geq \frac{m}{3e^{2B} + 1} \sinh^2 \left(\frac{\lambda}{4} \right).$$

VII. PROOF OUTLINE OF THEOREM 1

For the class $\mathcal{G}_{d,p}$, we pick the set of edges in the model which best models the data, given the freedom to pick the edge parameters as we see fit so long as they are bigger than λ . If we locate regions that could arise from each set of edges in the space of covariances $\{EX_s X_t\}$, then the separation between the regions determines the precision to which the covariances must be estimated in order to retrieve the edges accurately.

Suppose we have two models $\Lambda^{(1)}$ and $\Lambda^{(2)}$. The following Lemma shows that adjacent to any vertex s with different neighborhoods in the two models, there exists an edge (s, t) such that $|E_1 X_s X_t - E_2 X_s X_t|$ is suitably large.

Lemma 5. *Let $\Lambda^{(1)}$ be a model with edge set E_1 and $\Lambda^{(2)}$ be a model with edge set E_2 . For all edges $(s, t) \in (E_1 - E_2) \cup (E_2 - E_1)$*

$$\max_{u \in \{s, t\}, v \in V} |E_{\mathbb{P}_\lambda} X_u X_v - E_{\mathbb{P}_{\Lambda^{(2)}}} X_u X_v| \geq \frac{\sinh^2(\lambda/4)}{B(3e^{2B} + 1)}.$$

A Hoeffding-type large deviations result and an union bound yields the following result. If

$$n > \frac{(3e^{2B} + 1)^2}{\sinh^4(\frac{\lambda}{2})} \left(2 \log p + \log \frac{1}{\delta} \right)$$

then \exists decoder $q : X^n \rightarrow \mathcal{G}_{d,p}$ such that for all $\Lambda^{(i)}$, $1 \leq i \leq N$,

$$\mathbb{P}(q(X^n) \neq \Lambda^{(i)} | \Lambda^{(i)}) < \delta.$$

VIII. PROOF OUTLINE OF THEOREM 2

The trick is to find subsets of models in $\mathcal{G}_{d,p,\theta}$ (for any θ) that determine the necessary regions. Since we determine the necessary region for $\mathcal{G}_{d,p,\theta}$ no matter what the edge parameters θ are, the necessary region for $\mathcal{G}_{d,p}$ follows from the result on $\mathcal{G}_{d,p,\theta}$.

We require the following lemma to determine a lower bound on sample complexity.

Lemma 6. *The cardinality of $\mathcal{G}_{d,p,\theta}$ satisfies for $d \leq \frac{p}{2}$,*

$$\left(\left\lfloor \frac{p}{d+1} \right\rfloor! \right)^{d(d+1)/2} \leq |\mathcal{G}_{d,p,\theta}| \leq pd \binom{\frac{p}{2}}{\frac{pd}{4}},$$

therefore

$$\log |\mathcal{G}_{d,p,\theta}| = \Theta \left(pd \log \frac{p}{d} \right).$$

Proof For the upper bound on $|\mathcal{G}_{d,p,\theta}|$, observe that every model in $\mathcal{G}_{d,p,\theta}$ has at most $\frac{pd}{2}$ edges. An upper bound on $|\mathcal{G}_{d,p,\theta}|$ is provided by the number of graphs with at most $pd/2$ edges with no restrictions on degrees. To upper bound the later, note that the number of graphs with exactly r edges is $\binom{\frac{p}{2}}{r}$ and for $d \leq p/2$, the number of graphs with $pd/2$ edges is greater than the number of graphs with r edges for all $r \leq pd/2$, the upper bound on $\mathcal{G}_{d,p,\theta}$ follows.

For the lower bound, we proceed as follows. Group p vertices into $d+1$ even groups (throw away any remaining vertices). Pick a permutation of $\lfloor p/(d+1) \rfloor$, and form an bijection from group 1 to group 2 corresponding to the permutation. Similarly form an injection from group 1 to 3, \dots , $d+1$ using $d-1$ other permutations (we use up d permutations in all).

Similarly use $d-1$ permutations to connect from group 2 to groups 3, \dots , $d+1$; $d+1-i$ permutations to connect from group i to groups $i+1, \dots, d+1$ and so on.

Therefore each choice $1 + 2 + \dots + d = d(d+1)/2$ permutations leads to a different graph satisfying degree bound d . The lemma follows. \square

We reproduce the statement of Theorem 2 for convenience here: for all decoders $q : X^n \rightarrow \mathcal{G}_{d,p}$, if $B \geq 1$ and

$$n \leq \frac{1}{2} \frac{e^{B/2} d \lambda (\log pd - 1)}{32(e^\lambda - e^{-\lambda})} + \frac{1}{4} d \log \frac{p}{d}.$$

there \exists model $\Lambda^{(i)}$ such that

$$\mathbb{P}(q(X^n) \neq \Lambda^{(i)} | \Lambda^{(i)}) \geq \frac{1}{2}.$$

Proof outline We first consider the necessary region for the class $\mathcal{G}_{k,p,\theta}$. Let $k+1 \geq \binom{l}{2}$ for some l , consider the following two models: (i) $\Lambda^{(1)}$: the completely connected graph with an edge removed, say (s, t) and (ii) $\Lambda^{(2)}$:

the completely connected graph with a different edge removed, say (s', t') . Note that

$$\frac{\mathbb{P}_2(X_s X_t = 1)}{\mathbb{P}_2(X_s X_t = -1)} \leq \frac{\mathbb{P}_1(X_s X_t = 1)}{\mathbb{P}_1(X_s X_t = -1)} e^{2\lambda},$$

we obtain as in (6) that

$$\mathcal{J}(1, 2) \leq \frac{2(e^{2\lambda} - 1)}{\frac{\mathbb{P}_1(X_s X_t = 1)}{\mathbb{P}_1(X_s X_t = -1)}}. \quad (7)$$

The FKG inequality implies that reducing the parameter on every edge of $\Lambda^{(1)}$ to λ only reduces the correlation on the edge (s, t) . Using Lemma 2 and (7) that long as $B > 1$,

$$\mathcal{J}(1, 2) \leq \frac{2B \sinh(\lambda)}{e^{B/2}}.$$

We now consider the set of $k + 1$ models, each with a different edge removed from the complete graph. From Lemma 4, if

$$n < \frac{e^{B/2} \log k}{2B \sinh(\lambda)} \leq \frac{\log k}{\mathcal{J}(1, 2)},$$

then the error probability is $\geq \frac{1}{2}$. Similarly considering the set of models obtained with the addition of any edge to any one of the models above, say 1, (except for the one that completes the fully connected graph) yields that for the error probability to be $< \frac{1}{2}$,

$$n > \frac{2 \log p}{(e^\lambda - e^{-\lambda})}.$$

For the bounded degree case, we group the p vertices into sets of $d + 1$ vertices, and consider the graph G_0 obtained by fully connecting each subset of $d + 1$ vertices ($\lfloor p/(d + 1) \rfloor$ cliques of size $d + 1$). Consider the following subset of models: from G_0 , remove one edge. If $p \geq 2(d + 1)$, it follows that we obtain $\lfloor p/(d + 1) \rfloor \binom{d+1}{2} \geq pd/4$ such models, and for each distinct pair of models i and j so obtained

$$\mathcal{J}(i, j) \leq \frac{4B(e^\lambda - e^{-\lambda})}{e^B}.$$

Fano's inequality yields

$$n_- > \frac{\log pd - 1}{\mathcal{J}(1, 2)} \geq \frac{e^B(\log pd - 1)}{4B(e^\lambda - e^{-\lambda})} \geq \frac{e^{B/2} d \lambda (\log pd - 1)}{32(e^\lambda - e^{-\lambda})}.$$

Next, we show that $> d \log p$ are always required. Observe that for all n

$$I(X^n; \Lambda) \leq H(X^n) \leq np.$$

From Lemma 6, for all n ,

$$\frac{I(X^n; \Lambda)}{\log |\mathcal{G}_{d,p,\theta}|} \leq \frac{np}{pd \log \frac{p}{d}} \leq \frac{n}{d \log \frac{p}{d}}.$$

The theorem follows. \square

Acknowledgements

This work was partially supported by NSF grants CAREER-0545862 and DMS-0528488.

REFERENCES

- [1] R. J. Baxter. *Exactly solved models in statistical mechanics*. Academic Press, New York, 1982.
- [2] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, 48(3):259–279, 1986.
- [3] G. Bresler, E. Mossel, and A. Sly. Reconstruction of markov random fields from samples: Some easy observations and algorithms. In <http://arxiv.org/find/all/1/au:+bresler/0/1/0/all/0/1>.
- [4] D. Chickering. Learning Bayesian networks is NP-complete. *Proceedings of AI and Statistics*, 1995.
- [5] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Info. Theory*, IT-14:462–467, 1968.
- [6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, editors. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, 1998.
- [7] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI*, 6:721–741, 1984.
- [8] D. Guo and T. Tanaka. Generic multiuser detection and statistical physics. In M. Honig, editor, *Advances in Multiuser Detection*. John Wiley and Sons, Inc. To appear.
- [9] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 2006. To appear.
- [10] M. Mezard and A. Montanari. *Information, physics and computation*. Draft of book in preparation, 2008. Draft from <http://www.stanford.edu/~montanar/BOOK/book.html>.
- [11] N.P. Santhanam and M.J. Wainwright. Information theoretic limits of graphical model selection in high dimensions. Available from <http://www.eecs.berkeley.edu/~prasadsn/index.html>.
- [12] N.P. Santhanam and M.J. Wainwright. Information-theoretic limits of graphical model selection in high dimensions. In *Proceedings of IEEE Symposium on Information Theory*, July 2008.
- [13] N. Sourlas. Spin glass models as error correcting codes. *Nature*, 339:692–695, June 1989.
- [14] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction and search*. MIT Press, 2000.
- [15] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical report, UC Berkeley, Department of Statistics, No. 649, September 2003.
- [16] M. J. Wainwright, P. Ravikumar, and J. Lafferty. High-dimensional graph selection using ℓ_1 -regularized logistic regression. In *NIPS Conference*, December 2006.
- [17] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.