Inferring graphical structures

Piyush Srivastava* Di Wang†

Abstract

We consider the problem of inferring the underlying graph using samples from a Markov random field defined on the graph. In particular, we consider the special but interesting case when the underlying graph comes from a distribution on sparse graphs.

We provide matching upper and lower bounds for the sample-complexity of learning the underlying graph of a hard-core model, when the underlying graph model in $\mathcal{G}(n, d/n)$. We also survey some recent results on inferring pairwise Markov random fields from samples on graphs drawn from $\mathcal{G}(n, d/n)$ and from random graphs of degree at most d.

1 Introduction

1.1 Background

Markov random fields (MRFs) originated in statistical mechanics in the form of *spin systems* such as the Ising and Potts models [4, 6]. Since then, they have become an important tool for the modeling of interactions between subsystems in a wide variety of machine learning applications (see, e.g., [7] for a survey). In a typical application, a system is modeled as a network G of vertices, and a *configuration* of the system is an assignment of values from some fixed domain Ω to the vertices. The system is probabilistic, and the edges of the network determine the dependency structure between the vertices in the following sense. For any subsets S, T and U of nodes such that the removal of T disconnects S and U in G, the distribution of the configurations restricted to S and U are conditionally independent given the configuration on T. The celebrated Hammersley-Clifford theorem [3] guarantees that when G is finite, joint distributions on the nodes of G which satisfy the above conditional independence criterion are precisely those which factorize over the cliques in G. An important sub-class of MRFs is that of *pairwise* MRFs; these are MRFs in which the Hammersley-Clifford factorization has non-trivial contributions only from the vertices and edges in G (i.e., only from cliques of size at most 2). In this article, we will be concerned almost exclusively with these models.

One can associate two natural classes of inference problems with MRFs. Firstly, one can assume that the network G itself is known, in which case the natural inference problem is to use samples from the MRF to estimate the vertex and edge potentials that define the MRF (see section 2.1 for the formal definition of an MRF in terms of these potentials). However, in this article, we will be concerned with the so called *model selection* problem, which is concerned with inferring the edges in the MRF given samples from the MRF. More formally, we assume that the graph G on n nodes underlying the MRF is drawn from a known distribution (we will deal mostly with the Erdős-Rényi distribution $\mathcal{G}(n, d/n)$, and with the uniform distribution $\mathcal{G}_d(n)$ on graphs of degree at most d). We are then provided with $s(n)$ samples from the MRF, and our goal is to produce a graph $\hat{G}$ such that $G = \hat{G}$ with probability at least $1 - o_n(1)$ over the choice of G, the samples, and any randomization in our inference algorithm. The complexity of this problem is evaluated in terms of two parameters:

*Email: piyushsr@cs.berkeley.edu.
†Email: wangd@cs.berkeley.edu.

- The *sample complexity*, which is the smallest number of samples required in order to infer the underlying graph with good confidence.
- The *time complexity* of the inference algorithm itself.

1.2 Contributions

In this article, we consider the problem of inferring graphical structures when the underlying graph is drawn from the Erdős-Rényi distribution $\mathcal{G}(n, d/n)$. Specifically, we show upper and lower bounds on the sampling complexity for inferring the underlying graph G drawn from $\mathcal{G}(n, d/n)$ given samples from the hard-core model on G (this is a pairwise MRF which is a generalization of uniformly random independent sets: see section 2.2 for the formal definition). We also give a simple $n^{2+O(1/\log\log n)}$ algorithm for the reconstruction. Our results are summarized in the following theorem.

Theorem 1.1 (Informal)**.** *Consider a graph G drawn from $\mathcal{G}(n, d/n)$. There exists an algorithm which, given $n^{O(1/\log\log n)}$ samples from the hard core model on G, produces in time $n^{2+O(1/\log\log n)}$ a graph $\hat{G}$ such that $G = \hat{G}$ with probability at least $1 - o(1)$ over the random choice of G and the samples. Further, $n^{\Omega(1/\log\log n)}$ samples are required in order to achieve a probability of error smaller than $1/3$.*

The main technique used in inferring MRFs on bounded degree graphs (such as those drawn from $\mathcal{G}_d(n)$, as considered by Bresler, Mossel and Sly [2]) can be described as *neighborhood analysis*. In this method, the empirical correlations of a vertex v with all of its $n^{\Theta(d)}$ candidate neighborhoods are tested in order to infer its actual neighborhood. This technique is quite powerful, and provides a sample complexity of $O(\log n)$ with only mild restrictions on the potentials of the MRF. However, the running time of such an algorithm is $n^{\Omega(d)}$, and hence a direct application to the case of $\mathcal{G}(n, d/n)$ would yield a quasi-polynomial time algorithm (since $G(n, d/n)$ contains several vertices of degree $\Theta\left(\frac{\log n}{\log\log n}\right)$ with high probability). Anandkumar, Tan, Huang and Willsky [1] get around this difficulty with the additional assumption of correlation decay (see section 5 for more details). The results of both Bresler *et al.* [2] and Anandkumar *et al.* [1] are focused towards "soft-constraint" models, and provide tight $\Theta(\log n)$ upper and lower bounds on the sample complexity.

In contrast, in our setting of the hard core model (Theorem 1.1), we are able to leverage the "hard" constraints to provide a very simple reconstruction algorithm for $\mathcal{G}(n, d/n)$ which does not depend upon decay of correlations. Further, unlike the soft constraint setting of Bresler *et al.* [2] and Anandkumar *et al.* [1] where the right answer for the sample complexity turns out to be $\Theta(\log n)$, we are able to show that the sampling complexity is $n^{\Theta(1/\log\log n)}$. However, the fact that our algorithm depends crucially on the fact that our MRF has "hard-constraints" can be seen as a drawback of our Theorem 1.1 as compared to the results for soft-constraints models cited above.

We prove Theorem 1.1 in two parts; the upper bound is proved as Theorem 3.1 in section 3.1 and the lower bound on the sample complexity is proved as Theorem 3.2 in section 3.2. We then discuss some previous results on inferring MRFs on bounded degree graphs drawn from $\mathcal{G}_d(n)$ [2], and on the use of correlation decay to inference on graphs drawn from $\mathcal{G}(n, d/n)$ [1] in sections 4 and 5 respectively.

2 Preliminaries

We will denote by $\mathcal{G}(n, d/n)$ the Erdős-Rényi random graph model where each edge is present independently with probability d/n, and by $\mathcal{G}_d(n)$ the set of graphs of degree at most d on n vertices. With a slight abuse of notation, we will also sometimes denote the uniform distribution on $\mathcal{G}_d(n)$ by $\mathcal{G}_d(n)$. All graphs in the this article are assumed to be undirected.

2.1 Markov random fields

In this paper, we will mostly be concerned with the special class of pairwise Markov random fields (also known as spin systems). Given a finite set Ω of *spins*, and a graph $G = (V, E)$, such a Markov random field (MRF) is defined in terms of *edge activities* $w_{\{u,v\}} : \Omega^2 \to \mathbb{R}^+$ for each $e = \{u, v\} \in E$, and *vertex activities* $w_v : \Omega \to \mathbb{R}^+$ (the edge potentials are assumed to be symmetric functions of their argument). A *configuration* $\sigma \in \Omega^V$ is an assignment of spins to the vertices. The Markov random field then assigns a probability $p(\sigma)$ to each configuration σ where

$$p(\sigma) \propto w(\sigma) := \prod_{\{u,v\} \in E} w_{\{u,v\}}(\sigma_u, \sigma_v) \prod_{v \in V} w_v(\sigma_v).$$

A Markov random field is said to have *hard constraints* if there are configurations in Ω^V to which it assigns a probability of 0.

Example 2.1. (**Hard core model**). This is an MRF with hard constraints in which the allowed configurations are the independent sets of the graphs. Formally, we have $w_e(1,1) = 0$ and $w_e(0,1) = w_e(1,0) = w_e(0,0) = 1$ for all edges e, while $w_v(1) = \lambda$ and $w_v(0) = 1$ for all vertices v, where $\lambda > 0$ is also called the *fugacity*. Thus, for an independent set I, the probability $p(I)$ is proportional to $\lambda^{|I|}$; and the setting $\lambda = 1$ corresponds to the uniform distribution on all independent sets.

Example 2.2. (**(Zero field) Ising model**) This is a soft-constraint model with the set of spins $\Omega = \{+, -\}$, $w_v(+) = w_v(-) = 1$ for all vertices $v \in V$ and $w_e(+,+) = w_e(-,-) = e^\beta$ and $w_e(+,-) = w_e(-,+) = 1$ for all $e \in E$. The model was initially proposed in statistical physics as a tool for the qualitative study of phase transition in magnets. The parameter β is called the *inverse temperature*. The setting $\beta > 0$ is called *ferromagnetic*, since neighboring spins have a propensity for having *equal* spins; while the setting $\beta < 0$ is called *anti-ferromagnetic*, since neighboring spins have a propensity for having *different* spins.

2.2 Hard core model

We will need some simple facts about the marginals of the hard core model on trees and graphs. Let T be a tree rooted at a vertex v, and let v_i for $1 \le i \le d$ be the children of v. By a slight abuse of notation, we will denote by p_u the probability of vertex u in the tree T being occupied in the hard core distribution *restricted* to the subtree rooted at u. We call p_u the *occupation probability*. We further denote by R_v the *occupation ratio* $R_v := \frac{p_v}{1-p_v}$. The R_v's follow the well known recurrence (see, e.g., [8])

$$R_v = \lambda \prod_{i=1}^{d} \frac{1}{1 + R_{v_i}}. \tag{1}$$

Using eq. (1) (and, for example, Weitz's translation for marginals of the hard core on general graphs to trees), we obtain the following simple lower bound on the occupation probability in the hard core model.

Observation 2.1. *Consider the hard core model with fugacity λ on a graph G, and let v be a vertex in G of degree d. Then there exits a constant $c = c(\lambda)$ such that for a random independent set I chosen according to the hard core distribution,*

$$\mathbf{P}[v \in I] \ge \exp(-cd).$$

Further, if u is any other vertex such that $\{u, v\}$ is not an edge in G, then we also have

$$\mathbf{P}[v \in I \mid u \in I] \ge \exp(-cd).$$

3 Inference in the hard-core model

We now consider the problem of inferring the underlying graph from samples drawn from the hard core model. We assume that the underlying graph is drawn from the random graph model $\mathcal{G}(n, d/n)$.

3.1 Upper bound

We first give a simple algorithm for reconstructing the underlying graph using samples from the hard core model. Our algorithm obtains s samples $\boldsymbol{I} = (I_1, I_2, \ldots, I_s)$ from the hard core model on a graph G drawn from $\mathcal{G}(n, d/n)$, and then produces an output graph $\hat{G}$ as follows:

- For each candidate edge $\{u, v\}$:
  - If there exists an i, $1 \le i \le s$ such that one of the independent sets I_i contains both u and v, then reject $\{u, v\}$, else include $\{u, v\}$ in $\hat{G}$.
- Return $\hat{G}$.

We then prove the following theorem.

Theorem 3.1. *Consider the hard core model with fugacity $\lambda \in [\lambda_l, \lambda_u]$, where λ_l and λ_u are fixed constants. There exists a constant $c = c(\lambda_l, \lambda_u, d)$ such that for $s \ge \exp\left(c \cdot \frac{\log n}{\log\log n}\right)$, the output $\hat{G}$ of the above algorithm is the same as G with probability at least $1 - 1/n$. Here, the probability is over the random choice of G and of the random samples from the hard core model on G.*

Remark 3.1. Although Theorem 3.1 is stated and proved in the case when the fugacity λ is the same across all vertices, it is easy to see that the proof generalizes easily to the case where the fugacity is allowed to vary across vertices.

Theorem 3.1 shows that $n^{O(1/\log\log n)}$ samples are sufficient to reconstruct the underlying graph of the hard core model: we provide a matching lower bound of $n^{\Theta(1/\log\log n)}$ on the sample complexity in section 3.2. We also notice that the total time complexity of our algorithm is $O(ns) = O(n^{2+O(1/\log\log n)})$. We now prove Theorem 3.1.

Proof. Let $(1+\delta) := \frac{5}{d} \cdot \frac{\log n}{\log\log n}$. We will show that with high probability the degree d_v of every vertex in G is at most $d(1+\delta) = 5\log n/\log\log n$. By a standard Chernoff bound, we have

$$\begin{aligned}\mathbf{P}\left[d_v \ge d(1+\delta)\right] &\le \exp\left[-d\left((1+\delta)\log(1+\delta) - \delta\right)\right] \\ &\le \frac{1}{n^{4.9}}, \text{ for } n \text{ large enough.}\end{aligned}$$

Let $\mathcal{E}$ be the event that $d_v \le d(1+\delta)$ for all vertices in v. By a union bound, we then have $\mathbf{P}[\mathcal{E}] \ge 1 - n^{-3.9}$.

Now consider any pair of vertices u, v in G that are not connected by an edge. Consider a random independent set I drawn from the hard core model on G. From Observation 2.1, we then have

$$\mathbf{P}\left[u, v \in I \mid \mathcal{E}\right] \ge p := \exp\left(-2c \cdot \frac{5\log n}{\log\log n}\right),$$

where c is a constant depending only upon λ_l and λ_u. Now consider s independent samples $\boldsymbol{I} = (I_1, I_2, \ldots, I_s)$ from the hard core model on s. Using the definition of p, we see that

$$\mathbf{P}\left[\{u, v\} \not\subseteq I_i \text{for all } 1 \le i \le s \mid \mathcal{E}\right] \le (1-p)^s \le e^{-ps}.$$

Thus, by a union bound over the edges followed by correction for the condition of the high probability event $\mathscr{E}$, the probability of failure P_{failure} of the algorithm is given by

$$P_{\text{failure}} \leq n^2 e^{-ps} + \mathbf{P}\left[\neg \mathscr{E}\right],$$

which is at most $\frac{1}{n}$ for n large enough if we choose $s = \left\lceil \frac{5 \log n}{p} \right\rceil \leq \exp\left(3c \cdot \frac{5 \log n}{\log\log n}\right)$. □

3.2 Lower bound

We now proceed to prove a lower bound on the sample complexity for reconstruction in the hard core model. More precisely, we will show the following theorem on the performance of any deterministic maximum *a posteriori* (MAP) estimator for the underlying graph G. Since any other estimator has a probability of error lower bounded by the error of a MAP estimator, the result applies to all estimators.

Theorem 3.2. *Let G be a graph drawn from $\mathscr{G}(n, d/n)$. Consider a sequence $\boldsymbol{I} = (I_1, I_2, \ldots, I_s)$ of s independent samples from the hard core model with fugacity $\lambda > 0$ on G, and let* MAP$(\boldsymbol{I})$ *denote the output of the* MAP *estimator on input $\boldsymbol{I}$. Then, there exists a constant $c = c(\lambda, d)$ such that for $s \leq \exp\left(c \cdot \frac{\log n}{\log\log n}\right)$, we have*

$$\mathbf{P}_{G,\boldsymbol{I}}\left[\text{MAP}(\boldsymbol{I}) \neq G\right] \geq \frac{1}{2} - o_n(1).$$

Theorem 3.2 thus shows that at least $n^{\Theta(1/\log\log n)}$ samples are required to reconstruct the underlying graph of the hard core model. This complements Theorem 3.1 which showed that reconstruction is possible with $n^{O(1/\log\log n)}$ samples.

We now proceed to prove Theorem 3.2. We begin with the definition of an event S on which MAP fails with non-negligible probability. Let G be the underlying graph drawn from $\mathscr{G}(n, d/n)$ and let $\boldsymbol{I} = (I_1, I_2, \ldots, I_s)$ be a sequence of s samples drawn from the hardcore model on G. We then define the event S as follows:

$$S := \left\{(G, \boldsymbol{I}) \mid \exists G' \neq G \text{ with } \mathbf{P}\left[G\right] = \mathbf{P}\left[G'\right] \text{ and } \mathbf{P}\left[\bar{I} \mid G\right] = \mathbf{P}\left[\bar{I} \mid G'\right](1 + o_n(1))\right\}. \tag{2}$$

Notice that if $(G, \boldsymbol{I}) \in S$, then there is a graph $G' \neq G$ such that $\mathbf{P}\left[G \mid \boldsymbol{I}\right] = \mathbf{P}\left[G' \mid \boldsymbol{I}\right](1 + o_n(1))$. We therefore have the following simple observation.

Observation 3.3. $\mathbf{P}_{G,\boldsymbol{I}}\left[\text{MAP}\left(\boldsymbol{I}\right) \neq G \mid S\right] \geq \frac{1}{2} - o_n(1)$.

Thus, in order to establish Theorem 3.2, it is sufficient to show that there exists a constant $c = c(\lambda, d)$ such that for $s \leq \exp\left(c \cdot \frac{\log n}{\log\log n}\right)$, we have $\mathbf{P}\left[S\right] \geq 1 - o_n(1)$.

In order to prove this statement, we begin with the notion of a *faithful copy* of a small graph in another graph. Recall that a copy of a graph H in the complete graph K_n is a subgraph of K_n that is isomorphic to H [5].

Definition 3.1. (Faithful copy). Let $H = (V_H, E_H)$ be a fixed graph, and let $G = (V_G, E_G)$ be another graph. A copy T of H in the complete graph K_{V_G} is called a *faithful copy* of H in G if T is an induced subgraph of G such that *all* the edges in G incident on vertices in T are contained in T.

Note that our notion of having a faithful copy is slightly stronger than the usual notion of subgraph containment [5]. However, we can still prove similar results about existence of faithful copies as those known about subgraph containment.

Lemma 3.4. *There exists a constant $\alpha = \alpha(d)$ such that the following is true. Let $H_n = (V_n, E_n)$ be a sequence of forests such that $|V_n| \leq \alpha \cdot \frac{\log n}{\log\log n}$. Then, a graph G drawn from $\mathscr{G}(n, d/n)$ has a faithful copy of H_n with probability at least $1 - o_n(1)$.*

The proof of Lemma 3.4 requires some technical calculations, and we defer it to appendix A. We first show how to use it to prove Theorem 3.2.

Proof of Theorem 3.2. As discussed above, we only need to show that $\mathbf{P}[S] \geq 1 - o_n(1)$, where the event S is as defined in eq. (2).

Let G be drawn from $\mathcal{G}(n, d/n)$, and consider the graph H shown in Figure 1 (α is as in the statement of Lemma 3.4). Since H is a forest of size less that $\alpha \cdot \frac{\log n}{\log\log n}$, Lemma 3.4 implies that with probability at least $1 - o(1)$, G contains a faithful copy of H.

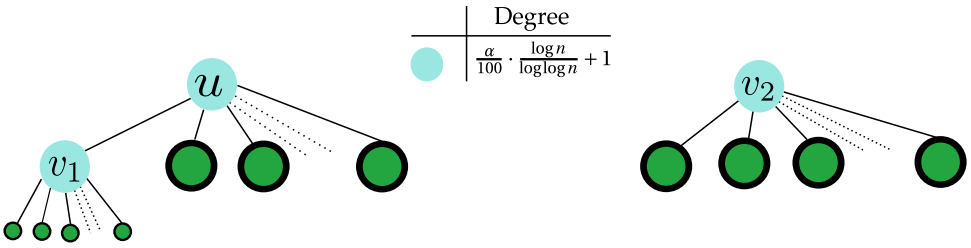

Figure 1: The subgraph H

We now condition on the event $\mathcal{E}_H$ that G contains a faithful copy of H, and define $G' = G - \{u, v_1\} + \{u, v_2\}$, where u, v_1 and v_2 are the special vertices in H identified in the figure. Since G and G' contain the same number of edges, we have $\mathbf{P}[G] = \mathbf{P}\left[G'\right]$. Further since the vertices u, v_1 and v_2, all have $\frac{\alpha}{100}\frac{\log n}{\log\log n}$ children, each of which, in turn, have zero children, in both G and G', the tree recurrence (eq. (1)) implies that for an independent set I drawn from the hard core model on either G or G', we have

$$\mathbf{P}[x \in I] \leq \exp\left(-c_1 \cdot \frac{\log n}{\log\log n}\right) \text{ for } x = u, v_1 \text{ or } v_2,$$

where $c_1 = c_1(\lambda, d)$ is a constant. Let $\mathcal{E}_{\text{same}}$ be the event that none of u, v_1, and v_2 are present in any of the s independently drawn samples $\boldsymbol{I} = (I_1, I_2, \ldots, I_s)$ from the hard core model on G. By a union bound, we therefore have $\mathbf{P}[\mathcal{E}_{\text{same}} \mid G] \geq 1 - a'$ and $\mathbf{P}\left[\mathcal{E}_{\text{same}} \mid G'\right] \geq 1 - a'$, where $a' := 3s \cdot \exp\left(-c_1 \cdot \frac{\log n}{\log\log n}\right)$.

Since conditioning on $\mathcal{E}_{\text{same}}$ makes the connectivity structure on the vertices u, v and w irrelevant, we further have $\mathbf{P}[I \mid \mathcal{E}_{\text{same}}, G] = \mathbf{P}[I \mid \mathcal{E}_{\text{same}}, G]$. Choosing $s = \exp\left(-c_1 \cdot \frac{\log n}{2\log\log n}\right)$, we therefore see that with probability at least $1 - a' = 1 - o_n(1)$, the observed vector $\boldsymbol{I}$ satisfies $\mathbf{P}[\boldsymbol{I} \mid G] = \mathbf{P}\left[\boldsymbol{I} \mid G'\right](1 + o_n(1))$.

Recalling that the above argument was conditioned on the event $\mathcal{E}_H$, and that $\mathbf{P}[\mathcal{E}_H] \geq 1 - o(1)$, we finally get that when the sample size s is chosen to be at most $\exp\left(-c_1 \cdot \frac{\log n}{2\log\log n}\right)$ as described above, $\mathbf{P}[S]$ is at lest $1 - o_n(1)$. As discussed above, this already proves the theorem. □

4 Inference in bounded degree graphs

In this section we briefly describe the results and proof techniques of Bresler *et al.* [2], who study the reconstruction problem for MRFs on graphs drawn from the uniform distribution on the set $\mathcal{G}_d(n)$ of graphs of degree at most d. They show that under mild conditions on the edge interactions of the underlying MRF (conditions which are true, e.g, for the Ising model), the sample complexity is $\Theta(\log n)$. They further give $n^{O(d)}$ reconstruction algorithms under these assumptions. We first describe their lower bound proof.

4.1 Lower Bound

The lower bound of $\Omega(d \log n)$ is demonstrated using a information-theoretic argument. The first observation is that it is sufficient to consider deterministic algorithms, since the smallest error probability is achieved by a deterministic maximum a posteriori (MAP) decision rule. Since the underlying distribution on $\mathcal{G}_d(n)$ is uniform, the lower bound proof then boils down to showing that at least $cd \log n$ samples are need for the size of the image of the MAP decision rule to be comparable to the size of $\mathcal{G}_d(n)$. This is achieved in the following lemma.

Lemma 4.1 ([2])**.** *Suppose $d \leq n^\alpha$ with $\alpha < 1$. Then the number of graphs with max degree at most d, $|\mathcal{G}_d(n)|$, satisfies*

$$\log |\mathcal{G}_d(n)| = \Omega(nd \log n).$$

It follows that we need $\Omega(d \log n)$ samples to reconstruct the graph.

4.2 Upper bound

As stated in the introduction, the upper bound proofs of Bresler *et al.* [2] are based on a "neighborhood analysis" technique. We present only a brief sketch of their argument here.

At a high level, the algorithms of Bresler *et al.* take the following form: given a vertex v, all possible candidate neighborhoods U are tested using empirical conditional probabilities. The test itself may be seen as looking for a witness w where w and v are non-trivially correlated conditioned on U. Under some mild conditions on the strength of the edge interactions, Bresler *et al.* are able to show that given $O(\log n)$ samples, the empirical estimates of these correlations are accurate enough (with high probability) that any U with $N(v) \nsubseteq U$ will fail the test.

However, testing for all candidate neighborhoods yields a $n^{O(d)}$ running time. Nevertheless, for models with correlation decay, Bresler *et al.* show that one only needs to consider candidate neighborhoods of v that are subsets of nodes which exhibit high empirical correlation with v. Further, under the assumption of correlation decay on a graph of degree at most some constant d, all such nodes are contained within a constant distance of v with high probability. Using this observation, Bresler *et al.* are able to improve the running time to $O(n^2 \log n)$ when correlation decay holds for the underlying MRF.

As pointed out above, the "neighborhood analysis" approach as outlined above will in general yield quasi-polynomial running time for models like $\mathcal{G}(n, d/n)$ which have high degree vertices. However, in the bounded degree setting, the results of Bresler *et al.* have the advantage of being applicable to a very wide range of MRFs.

5 Inference on $\mathcal{G}(n, d/n)$: correlation decay

Anandkumar *et al.* [1] study the graph reconstruction problem for several MRFs under the assumption of correlation decay. Of particular relevance to us here are their results for the Ising model on $\mathcal{G}(n, d/n)$. They show that under a suitable assumption about correlation decay and on the edge potentials, the sample complexity for reconstruction is $\Theta(\log n)$, with a polynomial running time for the reconstruction algorithm. Recall that a naïve application of the "neighborhood analysis" technique of Bresler *et al.* [2] gives a quasi-polynomial running time when the underlying graph model is $\mathcal{G}(n, d/n)$. We now give a brief sketch of the arguments of Anandkumar *et al.*.

The key observation for their upper bound proof is that graphs in $\mathcal{G}(n, d/n)$ are locally tree-like. In particular, the authors show that with high probability, for any two vertices u and v in a graph G drawn from $\mathcal{G}(n, d/n)$, there are at most two paths from u to v of length smaller than $\frac{\log n}{4 \log d}$. This implies that for any two nodes u, v not connected to each other by an edge, there exists a set of size 2

whose removal will eliminate all short paths between u and v. The assumption of correlation decay is then used to argue that the longer paths between u and v will make a negligible contribution to the correlation between the state and u and v. This leads to the $O(n^4)$ reconstruction algorithm where for each pair u, v of vertices, we look for two nodes x, y such that conditioned on x, y, the empirical correlation between u, v is insignificant. If such x, y exist, we deduce that there is no edge between u and v, otherwise we deduce that the edge $\{u, v\}$ is present in G. With the assumption of correlation decay (and conditions analogous to those of Bresler *et al.* [2] imposed on the edge potentials), it turns out that $O(\log n)$ samples are enough to estimate the required empirical correlations with high enough accuracy and confidence.

The authors also prove a $\Omega(\log n)$ lower bound for the sample complexity of the reconstruction problems they consider. Their proof is similar in spirit to the lower bound proof of Bresler *et al.* discussed in section 4.1. Recall that in the proof of Bresler *et al.*, it was argued that for a given number of samples $s = O(\log n)$, the range R of a deterministic estimator is of *size* at most $|\Omega|^{ns}$ which is much smaller than the number of graphs in $\mathcal{G}_d(n)$. The only difference in this setting comes from the fact that $\mathcal{G}(n, d/n)$ is not a uniform distribution on all the graphs in its domain. Hence, instead of comparing the size of R to the total number of graphs on which $\mathcal{G}(n, d/n)$ is supported, one has to argue that total probability mass of *any* set of at most $|\Omega|^{ns}$ graphs is negligible.

6 Future Work

An obvious next step is to generalize the results obtained here for the hard core model on $\mathcal{G}(n, d/n)$ to other hard constraint models such as random colorings; we expect this to be a straightforward extension of the methods used in the proofs in section 3. In contrast, a more challenging open problem is to simplify the correlation-decay based algorithms of Anandkumar *et al.* for inference on $\mathcal{G}(n, d/n)$. Currently, their algorithm may be viewed as doing a somewhat more global version of "neighborhood analysis" where all possible candidate small separators between two given vertices u and v are tested in order to infer whether u and v are connected by an edge in the graph. This makes the runtime of the reconstruction algorithm exponential in the size of such separators, and it will be interesting to find an algorithm which has a runtime of $n^{2+o(1)}$. We note here that our current methods do not seem to be of much help for this problem since they use the "hard-constraint" nature of the hard core distribution crucially.

Another avenue of future work is to extend Theorem 1.1 to other models of random graphs that are sparse on average, such as power law graphs.

Acknowledgments We thank Elchannan Mossel for several helpful discussions. We also thank Mohammad Moharrami for helpful comments.

References

[1] Anandkumar, A., Tan, V. Y. F., Huang, F., and Willsky, A. S. High-dimensional structure estimation in Ising models: Local separation criterion. *The Annals of Statistics 40*, 3 (June 2012), 1346–1375.

[2] Bresler, G., Mossel, E., and Sly, A. Reconstruction of markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques* (2008), A. Goel, K. Jansen, J. Rolim, and R. Rubinfeld, Eds., vol. 5171 of *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, pp. 343–356.

[3] HAMMERSLEY, J. M., AND CLIFFORD, P. Markov fields on finite graphs and lattices. Unpublished manuscript. Available at http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf., 1971.

[4] ISING, E. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift fur Physik 31* (Feb. 1925), 253–258.

[5] JANSON, S., ŁUCZACK, T., AND RUCINSKI, A. *Random Graphs*. Wiley Interscience, 2000.

[6] POTTS, R. B. Some generalized order-disorder transformations. *Proc. Cambridge Philos. Soc. 48* (1952), 106–109.

[7] WAINWRIGHT, M. J., AND JORDAN, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*, 1-2 (2008), 1–305.

[8] WEITZ, D. Counting independent sets up to the tree threshold. In *Proceedings of the 38th Annual ACM Symposium on the Theory of Computing* (2006), ACM, pp. 140–149.

A Proof of Lemma 3.4

Our proof has the same structure as the well-known second moment arguments for small-subgraph containment [5]. For a graph A, we will denote by e_A, v_A and a_A the number of edges, vertices and automorphisms of A. We have the following simple observation.

Observation A.1. *Let T be a copy of a graph A in the complete graph K_n. We then have*

$$\begin{aligned}\mathbf{P}\left[T \text{ is a faithful copy of } A \text{ in } \mathcal{G}(n,d/n)\right] &= \left(\frac{d}{n}\right)^{e_A}\left(1-\frac{d}{n}\right)^{v_A(n-v_A)+\binom{v_A}{2}-e_A}\\ &= \left(\frac{d}{n}\right)^{e_A} e^{-dv_A}(1+o_n(1)) \text{ when } v_A = o_n(\log n).\end{aligned}$$

Let G be a graph drawn from $\mathcal{G}(n,d/n)$. For any copy T of the forest H in the complete graph K_n, we will denote by I_T the indicator random variable for the event "T is a faithful copy of H in G", and by $X := \sum I_T$ the number of faithful copies of H in G. We will use the following version of Chebyshev's inequality:

$$\mathbf{P}[X=0] \le \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2}.$$

Thus, in order to prove Lemma 3.4, it is sufficient to prove that $\mathbf{Var}[X] \le \mathbf{E}[X]^2 o_n(1)$. We now begin with the proof.

Proof of Lemma 3.4. We begin by calculating $\mathbf{E}[X]$. Since there are $\binom{n}{v_H}\frac{v_H!}{a_H}$ distinct copies of H in the complete graph, Observation A.1 gives

$$\mathbf{E}[X] = \binom{n}{v_H}\frac{v_H!}{a_H}\left(\frac{d}{n}\right)^{e_H} e^{-dv_A}(1+o_n(1)) \ge \frac{1}{v_H!}n^{v_H-e_H}d^{e_H}e^{-dv_A}(1+o_n(1)).$$

We now consider the variance $\mathbf{Var}[X] = \sum_{S,T}\mathbf{E}[I_S I_T] - \mathbf{E}[I_S]I_T$, where the sum ranges over all pairs of copies (S,T) of H in K_n. We now break the sum into contributions coming from pairs (S,T) such that $S\cap T = U$, where U is some subgraph of H. We now use the fact that if $S\cap T = \emptyset$, then $\mathbf{E}[I_S I_T] = \mathbf{E}[I_S]\mathbf{E}[I_T](1+o_n(1))$, so that

$$\left(\sum_{S\cap T=\emptyset}\mathbf{E}[I_S I_T]\right) - \mathbf{E}[X]^2 \le \mathbf{E}[X]^2 o_n(1). \tag{3}$$

We now consider pairs such that $S \cap T$ is a copy of U, where U is some subgraph of H with $v_U > 0$. We then have

$$\sum_{S \cap T = U} \mathbf{E}[I_S I_T] \leq n^{2v_H - v_U - 2e_H + e_U} d^{2e_H - e_U} e^{-d(2v_H - v_U)} (1 + o_n(1)). \tag{4}$$

We now denote by W the subgraph of H with $v_W > 0$ for which the quantity in eq. (4) is maximized. Since there are at most 2^{e_H} subgraphs of H, we can combine eqs. (3) and (4) to get

$$\frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2} \leq o_n(1) + \frac{2^{e_H}(v_H!)^2}{n^{v_w - e_W}} d^{-e_W} e^{de_W} (1 + o_n(1)). \tag{5}$$

Since W is a subgraph of a forest H with $v_W > 0$, we must have $v_w - e_w \geq 1$. Further, we can choose α a small enough constant so that when $v_H \leq \alpha \cdot \frac{\log}{\log\log n}$, we have $v_H! = O(n^{0.1})$. Substituting these into eq. (5), we get the required bound $\frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2} \leq o_n(1)$. □