Learning graphical models from the Glauber dynamics

Learning graphical models from the Glauber dynamics

Guy Bresler[1] David Gamarnik[2] Devavrat Shah[1]

Laboratory for Information and Decision Systems
Department of Electrical Engineering and Computer Science[1]
Operations Research Center and Sloan School of Management[2]
Massachusetts Institute of Technology
{gbresler,gamarnik,devavrat}@mit.edu

***Abstract*— In this paper we consider the problem of learning undirected graphical models from data generated according to the Glauber dynamics. The Glauber dynamics is a Markov chain that sequentially updates individual nodes (variables) in a graphical model and it is frequently used to sample from the stationary distribution (to which it converges given sufficient time). Additionally, the Glauber dynamics is a natural dynamical model in a variety of settings. This work deviates from the standard formulation of graphical model learning in the literature, where one assumes access to i.i.d. samples from the distribution.**

Much of the research on graphical model learning has been directed towards finding algorithms with low computational cost. As the main result of this work, we establish that the problem of reconstructing binary pairwise graphical models is *computationally tractable* when we observe the Glauber dynamics. Specifically, we show that a binary pairwise graphical model on p nodes with maximum degree d can be learned in time $f(d)p^2 \log p$, for a function $f(d)$, using nearly the information-theoretic minimum number of samples.

I. Introduction

Examples of data one might usefully model as being generated according to a Markov process include the dynamics of agents in a coordination game, the fluctuations of stocks or other financial data, behavior of users in a social network, and spike patterns in neural networks.

The focus of this paper is on learning the nature of Markovian dynamics from observed data governed by local interactions. Concretely, we suppose that such local interactions are represented by a graphical model. We observe a single-site dynamics, specifically the so-called Glauber dynamics, and wish to learn the graph underlying the model.

This work fits within a broader theme of learning graphical models from data, a problem traditionally posed assuming access to i.i.d. samples from the model. While the assumption of i.i.d. samples makes sense as an abstraction (as well as in some practical scenarios), observations in many settings are correlated over time and in this case it is more natural to assume that samples are generated according to a Markov process. In general the distribution of such samples can be far from i.i.d.

The problems of learning and of generating samples are known to be related. On one hand, learning graphical models from i.i.d. samples is algorithmically challenging [1], [2], [3], and on the other hand, generating samples from distributions represented by graphical models is hard in general [4]. In the literature, much work has focused on trying to find low-complexity algorithms, both for learning as well as for generating samples, under various restrictions to the graphical model. Interestingly, related conditions (such as spatial and temporal mixing) have turned out to be central to most approaches.

Learning graphical models from i.i.d. samples appears to be challenging when there are correlations between variables on a global scale, as this seems to require a global procedure. Our results show that observing a *local* process allows to learn distributions with global correlations by *temporally* isolating the local structure.

A. Complexity of graphical model learning

A number of papers, including [5], [6], and [7] have suggested finding each node's neighborhood by exhaustively searching over candidate neighborhoods and checking conditional independence. For graphical models on p nodes of maximum degree d, such a search takes time (at least) on the order of p^d. As d grows, the computational cost becomes prohibitive, and much effort by the community has focused on trying to find algorithms with lower complexity.

Writing algorithm runtime in the form $f(d)p^{c(d)}$, for high-dimensional (large p) models the exponent $c(d)$ is of primary importance, and we will think of low-complexity algorithms as having an exponent $c(d)$ that is bounded by a constant independent of d.

Previous works proposing low-complexity algorithms either restrict the graph structure or the nature of the interactions between variables. The seminal paper of Chow and Liu [8] makes a model restriction of the first type, assuming that the graph is a tree; generalizations include to polytrees [9], hypertrees [10], tree mixtures [11], and others. Among the many possible assumptions of the second type, the correlation decay property (CDP) is dis-

tinguished: nearly all existing low-complexity algorithms require the CDP [3]. An exception is [12], which shows a family of antiferromagnetic models that can be learned with low complexity despite strongly violating the CDP.

Informally, a graphical model is said to have the correlation decay property (CDP) if any two variables σ_s and σ_t are asymptotically independent as the graph distance between s and t increases. The CDP is known to hold for a number of pairwise graphical models in the so-called high-temperature regime, including Ising, hard-core lattice gas, Potts (multinomial), and others (see the survey article [13] as well as, e.g., [14], [15], [16], [17], [18], [19], [20]).

It was first observed in [6] that it is possible to efficiently learn models with (exponential) decay of correlations, under the additional assumption that neighboring variables have correlation bounded away from zero. A variety of other papers including [21], [22], [23], [24] give alternative low-complexity algorithms, but also require the CDP. A number of structure learning algorithms are based on convex optimization, such as Ravikumar et al.'s [25] approach using regularized node-wise logistic regression. While this algorithm is shown to work under certain incoherence conditions and does not explicitly require the CDP, Bento and Montanari [3] showed through a careful analysis that the algorithm provably fails to learn ferromagnetic Ising models on simple families of graphs without the CDP. Other convex optimization-based algorithms such as [26], [27], [28] require similar incoherence or restricted isometry-type conditions that are difficult to interpret in terms of model parameters, and likely also require the CDP.

Most computationally efficient sampling algorithms (which happen to be based on the Markov Chain Monte Carlo method) require a notion of temporal mixing and this is closely related to spatial mixing or a version of the CDP (see, e.g., [29], [30], [31]). Thus, under a class of "mixing conditions", we can both generate (i.i.d.) samples efficiently as well as learn graphical models efficiently from such i.i.d. samples.

B. Main results

We give an algorithm that learns the graph structure underlying an arbitrary undirected binary pairwise graphical model from the Glauber dynamics, even without any mixing or correlation decay property. Concretely, in Theorem 2 we show that the algorithm learns the graph underlying any undirected binary pairwise graphical model over p nodes with maximum vertex degree d, given $\Omega(\log p)$ updates of the Glauber dynamics per node, starting from any initial state, with runtime $f(d)p^2 \log p$. The number of samples required by the algorithm is nearly information-theoretically optimal, as shown in the lower bound of Theorem 5.

C. Other related work

Several works have studied the problem of learning the graph underlying a random process for various processes. These include learning from epidemic cascades [32], [33], [34] and learning from delay measurements [35]. Another line of research asks to find the source of infection of an epidemic by observing the current state, where the graph is known [36], [37].

More broadly, a number of papers in the learning theory community have considered learning functions (or concepts) from examples generated by Markov chains, including [38], [39], [40], [41]. The present paper is similar in spirit to that of Bshouty et al. [40] showing that it is relatively easy to learn DNF formulas from examples generated according to a random walk as compared to i.i.d. samples.

The literature on the Glauber dynamics is enormous and we do not attempt to summarize it here. However, we remark that the Glauber dynamics is equivalent to a model of noisy coordination games and has been studied in that context by various authors: Saberi and Montanari [42] studied the impact of graph structure on rate of adoption of innovations, Berry and Subramanian [43] studied the problem of inferring the early adopters from an observation at a later time.

D. Outline

The rest of the paper is organized as follows. In Section II we define the model and formulate the learning problem. In Section III we present our structure learning algorithm and analysis. Then in Section IV we give an information-theoretic lower bound on the number of samples necessary in order to reconstruct with high probability.

II. Problem statement

A. Ising model.

We consider the Ising model on a graph $G = (V, E)$ with $|V| = p$. The notation ∂i is used to denote the set of neighbors of node i, and the degree $|\partial i|$ of each node i is assumed to be bounded by d. To each node $i \in V$ is associated a binary random variable (spin) σ_i. Each configuration of spins $\sigma \in \{-1, +1\}^V$ is assigned probability according to the Gibbs distribution

$$\mathsf{P}(\sigma) = \frac{1}{Z} \exp\left(\sum_{\{i,j\} \in E} \theta_{ij} \sigma_i \sigma_j \right). \tag{1}$$

Here Z is the partition function and serves to normalize the distribution. The distribution is parameterized by the vector of edge couplings $(\theta_{ij}) \in \mathbb{R}^E$, assumed to satisfy

$$\alpha \le |\theta_{ij}| \le \beta \quad \text{for } \{i,j\} \in E$$

for some constants $0 < \alpha \le \beta$. We can alternatively think of $\theta \in \mathbb{R}^{\binom{p}{2}}$, with $\theta_{ij} = 0$ if $\{i,j\} \notin E$. For a graph G, let

$$\Omega_{\alpha,\beta}(G) = \{\theta \in \mathbb{R}^{\binom{p}{2}} : \alpha \le |\theta_{ij}| \le \beta \text{ if } \{i,j\} \in E, \text{ and } \theta_{ij} = 0 \text{ otherwise}\}$$

be the set of parameter vectors corresponding to G.

The model (1) does not have node-wise parameters (that is, the external field is zero); while we restrict to this case for simplicity, similar results to those presented hold with suitable minor modifications to accommodate nonzero external fields.

The distribution specified in (1) is a *Markov random field*, and an implication is that each node is conditionally independent of all other nodes given the values of its neighbors. This allows to define a natural Markov chain known as the Glauber dynamics.

B. The Glauber dynamics.

The Glauber dynamics (also sometimes called the Gibbs Sampler) is a natural and well-studied reversible Markov chain defined for any Markov random field. For mathematical convenience we use both the continuous-time and discrete-time versions. We describe here the continuous-time dynamics, writing σ^t for the configuration at time $t \geq 0$. The process is started at some arbitrary (possibly random) initial configuration $\sigma^0 \in \{-1,+1\}^p$, and each node is updated at times given by an independent Poisson process of rate one. If spin σ_i is updated at time t, it takes on value $+1$ with probability

$$\mathsf{P}(\sigma_i = +1 \,|\sigma^t_{V\setminus\{i\}}) = \frac{\exp\left(2\sum_{j\in\partial i}\theta_{ij}\sigma^t_j\right)}{1+\exp\left(2\sum_{j\in\partial i}\theta_{ij}\sigma^t_j\right)}\,, \qquad (2)$$

and is -1 otherwise. Notably, each spin update depends only on neighboring spins. Equation (2) and the bounded coupling assumption $|\theta_{ij}| \leq \beta$ implies that for any $x \in \{-1,+1\}^{\partial i}$,

$$\min\{\mathsf{P}(\sigma_i = +1 \,|\sigma^t_{\partial i} = x), \mathsf{P}(\sigma_i = -1 \,|\sigma^t_{\partial i} = x)\} \geq \tfrac{1}{2}e^{-2\beta d}\,. \qquad (3)$$

This is a lower bound on the randomness in each spin update and will be used later.

The Glauber dynamics can be simulated efficiently for any bounded-degree undirected graphical model, and it is a plausible generating process for observed samples in various settings. One can check that the Gibbs distribution (1) is stationary for the Glauber dynamics. If the dynamics quickly approaches stationarity (that is, the mixing time is small), then it can be used to simulate i.i.d. samples from (1). But there are families of graphs for which any local Markov chain, including the Glauber dynamics, is known to converge exponentially slowly (see, e.g., [4]), and moreover the availability of i.i.d. samples violates conjectures in complexity theory in that it allows to approximate the partition function [44]. While it is difficult to imagine nature producing i.i.d. samples from such models, there is no such issue with the Glauber dynamics (or any other local Markov chain).

C. Graphical model learning

Our goal is to learn the graph $G = (V, E)$ underlying a graphical model of the form (1), given access to observations from the Glauber dynamics. We assume that the identity of nodes being updated is known; learning without this data is potentially much more challenging, because in that case information is obtained only when a spin flips sign, which may occur only in a small fraction of the updates.

For the purposes of recording the node update sequence it is more convenient to work with a discrete time (heat-bath) version of the chain, where each sample is taken immediately after a node is updated. In this case we denote the sequence of n samples by $\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(n)}$ and the corresponding node identities at which updates occur by $I^{(1)}, I^{(2)}, \ldots, I^{(n)}$. The value of $I^{(1)}$ is arbitrarily set to (say) one since the first configuration does not arise from a node update. The sequence of n samples is denoted by

$$X = (\sigma^{(l)}, I^{(l)})_{1\leq l\leq n} \qquad (4)$$

and is therefore an element of the product space

$$\mathcal{X} = (\{-1,+1\}^p)^n \times [p]^n\,.$$

We suppose that the continuous-time chain is observed for T units of time, so there are in expectation Tp spin updates. This number is tightly concentrated around the mean, and our arguments are not sensitive to a small amount of randomness in the number of samples n, so for convenience we deterministically set $n = Tp$.

As mentioned before, the underlying graph G is assumed to have maximum node degree bounded by d, and we denote the set of all such graphs on p nodes by $\mathcal{G}_{p,d}$. A *structure learning algorithm* is a (possibly randomized) map

$$\phi : (\{-1,+1\}^p)^n \times [p]^n \to \mathcal{G}_{p,d}\,.$$

The performance of a structure learning algorithm is measured using the zero-one loss, and the risk under some vector $\theta \in \Omega(G)$ of parameters corresponding to a graph $G \in \mathcal{G}_{p,d}$ is given by

$$\mathsf{P}_\theta(\phi(X) \neq G)\,.$$

The minimax risk is the best algorithm's worst-case risk (probability of error) over graphs and corresponding parameter vectors, namely

$$R_{p,d,n} \triangleq \min_{\phi} \max_{\substack{G\in\mathcal{G}_{p,d}\\ \theta\in\Omega(G)}} \mathsf{P}_\theta(\phi(X) \neq G)\,.$$

The basic questions we seek to address are what triples n, p, d result in the minimax risk $R_{p,d,n}$ tending to zero as these parameters tend to infinity, and can we find an efficient algorithm.

III. Structure learning algorithm

A. Idealized test

We determine the presence of edges in a decoupled manner, focusing on a single pair of nodes i and j. Our test is based on the identity (derived via Eq. (2))

$$e^{4\theta_{ij}} = \frac{p^+(1-p^-)}{p^-(1-p^+)}\,, \qquad (5)$$

where for an arbitrary assignment $x_{\partial i\setminus\{j\}}$ we define

$$p^+(x_{\partial i\setminus\{j\}}) = \mathsf{P}(\sigma_i = +1 | \sigma_{\partial i\setminus\{j\}} = x_{\partial i\setminus\{j\}}, \sigma_j = +1)$$
$$p^-(x_{\partial i\setminus\{j\}}) = \mathsf{P}(\sigma_i = +1 | \sigma_{\partial i\setminus\{j\}} = x_{\partial i\setminus\{j\}}, \sigma_j = -1)$$

(We will often leave implicit the dependence of p^+ and p^- on $x_{\partial i\setminus\{j\}}$.) The identity (5) holds whether or not the edge $\{i,j\}$ is present, since $\{i,j\} \notin E$ implies $\theta_{ij} = 0$, and this agrees with σ_i and σ_j being conditionally independent given $\sigma_{\partial i}$ (in which case $p^+ = p^-$).

Instead of attempting to estimate the right-hand side of (5) from samples, we claim that if our goal is merely to decide between $\theta_{ij} = 0$ and $|\theta_{ij}| \geq \alpha$, it suffices to estimate the much simpler quantity $p^+ - p^-$. This will be justified using the following bound.

Lemma 1: Let a and b be real numbers with $0 < a \leq b < 1$ and $a \leq \frac{1}{2}$. Then

$$b - a \leq \frac{b(1-a)}{a(1-b)} - 1 \leq \frac{b-a}{a(1-b)^2}\,.$$

Proof: Let $g(y) = \frac{y}{1-y}$. Then for $y \in (0,1)$, $g'(y) = (1-y)^{-2} > 1$. It follows that $g'(y) \leq (1-b)^{-2}$ for $y \in [a,b]$ and also from $a \leq \frac{1}{2}$ we get $a^{-1} \geq \frac{1-a}{a} \geq 1$. Combining these ingredients gives

$$\begin{aligned} b - a \leq g(b) - g(a) &\leq \frac{1-a}{a}\left(\frac{b}{1-b} - \frac{a}{1-a}\right) \\ &= \frac{b(1-a)}{a(1-b)} - 1 \leq \frac{1}{a}(g(b) - g(a)) \leq \frac{b-a}{a(1-b)^2}\,. \end{aligned}$$ ∎

Let us momentarily assume that $p^+ \geq p^-$ and $p^- \leq \frac{1}{2}$. The conditional probability lower bound (3) implies $\min\{1-p^+, p^-\} \geq \frac{1}{2}e^{-2\beta d}$. Lemma 1 together with (5) gives

$$p^+ - p^- \leq e^{4\theta_{ij}} - 1 \leq 8e^{8\beta d}(p^+ - p^-)\,.$$

The assumption $p^- \leq \frac{1}{2}$ is without loss of generality by replacing p^- and p^+ by $1-p^-$ and $1-p^+$, respectively. If $p^+ < p^-$, which happens if and only if $\theta_{ij} < 0$, we get a similar sequence of inequalities:

$$p^- - p^+ \leq e^{-4\theta_{ij}} - 1 \leq 8e^{8\beta d}(p^- - p^-)\,.$$

These can be combined to give

$$\operatorname{sign}(\theta_{ij})(p^+ - p^-) \leq e^{4|\theta_{ij}|} - 1 \leq \operatorname{sign}(\theta_{ij}) 8e^{8\beta d}(p^+ - p^-)\,. \tag{6}$$

We emphasize that this inequality holds for any assignment $x_{\partial i\setminus\{j\}}$.

It turns out to be possible to crudely estimate the quantity $p^+ - p^-$ in (6) to determine if it is equal to zero. It is important that the sign of $p^+ - p^-$ does not depend on the configuration $x_{\partial i\setminus\{j\}}$, as this allows to accumulate contributions from many samples. The following scenario gives intuition for why sequential updates allow to do this. Suppose that σ_i is updated, followed by σ_j flipping sign, followed by yet another update of σ_i, with no other spins updated. Since only σ_j has changed in between updates to σ_i, we can hope to get an estimate of the effect of σ_j on σ_i. To produce the sequence of events just described requires observing the process for $\Omega(p^2)$ time; we next show how to achieve a similar outcome sufficient for learning the structure, in time only $O(\log p)$.

B. Estimating edges

We define a few events to be used towards estimating the effect of an edge, as captured by $|p^+ - p^-|$. To this end, consider restriction of the process $\sigma^t \in \{-1,+1\}^p$ to an interval, written as

$$\sigma^{[t_1,t_2)} = (\sigma^t)_{t_1 \leq t < t_2}\,.$$

For a positive number L, let $A_{ij}(\sigma^{[0,L)})$ be the event that node i is selected at least once in the first $L/3$ time-units but not node j, node j is selected at least once in the second $L/3$ time-units but not node i, and node i is selected at least once in the final $L/3$ time-units but not node j. It is immediate from the Poisson update times that

$$\mathsf{P}(A_{ij}(\sigma^{[0,L)})) = [(1 - e^{-L/3})e^{-L/3})]^3 := q\,. \tag{7}$$

(We denote this quantity by q since it will be used often.) Next, define the event that σ_j is opposite at time $L/3$ versus $2L/3$,

$$B_{ij}(\sigma^{[0,L)}) = \{\sigma_j^{L/3} \neq \sigma_j^{2L/3}\}\,,$$

and take the intersection of the two events,

$$C_{ij}(\sigma^{[0,L)}) = A_{ij}(\sigma^{[0,L)}) \cap B_{ij}(\sigma^{[0,L)})\,.$$

Whenever σ_j is updated, by Equation (3) both the probabilities of flipping or staying the same are at least $\frac{1}{2}e^{-2d\beta}$, regardless of the states of its neighbors. It follows that the last update of σ_j in the interval $[\frac{L}{3}, \frac{2L}{3}]$ has at least probability $\frac{1}{2}e^{-2d\beta}$ of being opposite to $\sigma_j^{L/3}$, so

$$\mathsf{P}(C_{ij}) = \mathsf{P}(A_{ij}) \cdot \mathsf{P}(B_{ij} | A_{ij}) \geq \tfrac{1}{2}\mathsf{P}(A_{ij})e^{-2d\beta}\,. \tag{8}$$

Note that determining the occurrence of C_{ij} does not require knowing anything about the graph.

We now define the statistic that will be used to estimate presence of a given edge: For each $k \geq 1$ and $1 \leq i < j \leq p$, let

$$\begin{aligned} X_{ij}^{(k)} &= X_{ij}(\sigma^{[(k-1)L,kL)}) \\ &= \mathbb{1}_{C_{ij}(\sigma^{[(k-1)L,kL)})}(-1)^{\mathbb{1}\{\sigma_j^{L/3} = +1\}}(\sigma_i^{L/3} - \sigma_i^L)\,. \end{aligned}$$

The value $X_{ij}^{(k)} \in \{-2, 0, +2\}$ can be computed by an algorithm with access to the process $\sigma^{[(k-1)L,kL)}$. The idea is that $\mathsf{E}X_{ij}^{(k)}$ gives a rough estimate of the effect of spin j on spin i by counting the number of times σ_i has differing updates when σ_j has changed. It is necessary that few or no neighbors of i are updated during the time-interval, as these changes could overwhelm the effect due to σ_j. We will see later that choosing L sufficiently small ensures this is usually the case.

C. Structure learning algorithm

We now present the structure learning algorithm. In order to determine presence of edge $\{i,j\}$ the algorithm divides up time into intervals of length L, estimates $\mathsf{E}X_{ij}$ from the intervals, and compares $|\mathsf{E}X_{ij}|$ to a threshold τ.

Algorithm 1 GLAUBERLEARN$(\sigma^{[0,T)}, L, \tau)$

```
1: Let Ê = ∅ and k_max = ⌊T/L⌋.
2: For 1 ≤ i < j ≤ p
3:    If |(1/k_max) Σ_{k=1}^{k_max} X_ij^(k)| ≥ τ
4:    Then add edge {i,j} to Ê.
5: Output Ê.
```

Theorem 2: Consider the Ising model (1) on a graph G with maximum degree d and couplings bounded as $\alpha \le |\theta_{ij}| \le \beta$. Let $\sigma^{[0,T)}$ denote the continuous-time Glauber dynamics started from any configuration σ^0. If

$$L = \frac{\alpha}{16d} e^{-10d\beta}, \quad \tau = 3Ldq, \quad T \ge \frac{10^6 e^{20d\beta}}{\alpha^2} \log p\,,$$

where $q = \mathsf{P}(A_{ij}) = [(1-e^{-L/3})e^{-L/3})]^3$, then GLAUBERLEARN outputs the correct edge set with probability $1-\frac{1}{p}$ with runtime $O(p^2 \log p)$.

In the remainder of this section we work towards proving Theorem 2. We first bound the runtime of the algorithm. Suppose that when the samples are collected, they are stored as a list for each node giving the times the node is updated and the new value. Each computation in Line 3 takes time $O(\log p)$, and this is done for $O(p^2)$ pairs i,j, which gives the stated runtime.

Since the Glauber dynamics is time-homogeneous (and Markov), $\mathsf{E}(X_{ij}^{(k)}|\sigma^{(k-1)L} = x)$ does not depend on the index k. Hence, we use the shorthand $\mathsf{E}_x X_{ij}$ for $\mathsf{E}(X_{ij}^{(k)}|\sigma^{(k-1)L} = x)$ and similarly for $\mathsf{P}_x(\cdot) = \mathsf{P}(\cdot|\sigma^{(k-1)L} = x)$.

Let $D_{ij}(\sigma^{[t_1,t_2)})$ be the event that none of the neighbors of i, aside from possibly j, are selected in time-interval $[t_1,t_2)$. Since D_{ij} depends on disjoint Poisson clocks from those determining A_{ij}, the two events are independent (however, D_{ij} is not necessarily independent of B_{ij}). It is immediate, again from the Poisson times of updates, that

$$\begin{aligned}\mathsf{P}(D_{ij}(\sigma^{[(k-1)L,kL)})) &= \mathsf{P}(D_{ij}(\sigma^{[0,L)})) \\ &= \left(e^{-L}\right)^{|\partial i\setminus\{j\}|} \ge \left(e^{-L}\right)^d. \end{aligned} \tag{9}$$

At this point it is possible to make a connection to the idealized edge test formula (6). Conditioning on D_{ij}, our edge statistic has expectation

$$\begin{aligned}&\mathsf{E}_x(X_{ij}|D_{ij}) \\ &= \mathsf{E}_x(X_{ij}|C_{ij},D_{ij})\cdot \mathsf{P}_x(C_{ij}|D_{ij}) \\ &= \mathsf{E}_x\big((-1)^{\mathbb{1}\{\sigma_j^{L/3}=+1\}}(\sigma_i^{L/3}-\sigma_i^L)|C_{ij},D_{ij}\big)\cdot \mathsf{P}_x(C_{ij}|D_{ij}) \\ &= 2\big(\mathsf{P}(\sigma_i=+1|\sigma_{\partial i\setminus\{j\}}=x_{\partial i\setminus\{j\}},\sigma_j=+1) \\ &\quad - \mathsf{P}(\sigma_i=+1|\sigma_{\partial i\setminus\{j\}}=x_{\partial i\setminus\{j\}},\sigma_j=-1)\big)\cdot \mathsf{P}_x(C_{ij}|D_{ij}) \\ &= 2\big(p^+(x_{\partial i\setminus\{j\}})-p^-(x_{\partial i\setminus\{j\}})\big)\mathsf{P}_x(C_{ij}|D_{ij})\,. \end{aligned} \tag{10}$$

Of course, without knowing the neighbors of i it is not clear whether or not event D_{ij} has occurred, but as shown next in Lemma 3, if L is small enough, then D_{ij} occurs frequently and X_{ij} gives a good estimate.

Lemma 3: We have the following estimates:

(i) If $\{i,j\}\in E$, then for any $x\in\{-1,+1\}^p$,

$$\operatorname{sign}(\theta_{ij})\cdot \mathsf{E}_x X_{ij} \ge 2q\Big(|\theta_{ij}|\cdot \tfrac{1}{4}e^{-10d\beta}e^{-Ld} - Ld\Big)$$

(ii) If $\{i,j\}\notin E$, then for any $x\in\{-1,+1\}^p$,

$$|\mathsf{E}_x X_{ij}| \le 2qLd\,.$$

Proof: To begin, conditioning on D_{ij} gives

$$\mathsf{E}_x X_{ij} = \mathsf{E}_x(X_{ij}|D_{ij})\mathsf{P}_x(D_{ij}) + \mathsf{E}_x(X_{ij}|D_{ij}^c)\mathsf{P}_x(D_{ij}^c)\,. \tag{11}$$

In both cases (i) and (ii) we have

$$\begin{aligned}|\mathsf{E}_x(X_{ij}|D_{ij}^c)\mathsf{P}_x(D_{ij}^c)| &\overset{(a)}{\le} 2\mathsf{P}_x(C_{ij}|D_{ij}^c)\mathsf{P}_x(D_{ij}^c) \\ &\overset{(b)}{\le} 2\mathsf{P}_x(A_{ij}|D_{ij}^c)P_x(D_{ij}^c) \\ &\overset{(c)}{=} 2\mathsf{P}_x(A_{ij})P_x(D_{ij}^c) \\ &\overset{(d)}{\le} 2q(1-e^{-Ld}) \qquad (12)\\ &\overset{(e)}{\le} 2qLd\,. \qquad (13)\end{aligned}$$

Inequality (a) is by the crude estimate $|(-1)^{\mathbb{1}\{\sigma_j(L/3)=+1\}}(\sigma_i(L/3)-\sigma_i(L))| \le 2$, (b) follows from the containment $C_{ij}\subseteq A_{ij}$, (c) is by independence of A_{ij} and D_{ij}, (d) is obtained by plugging in (7) and (9), and (e) follows from the inequality $e^{-t}\ge 1-t$.

We first prove case (ii). If edge $\{i,j\}$ is not in the graph, then flipping only spin σ_j does not change the conditional distribution of spin σ_i, assuming the neighbors of i remain unchanged, and it follows from (10) that

$$\mathsf{E}_x(X_{ij}|D_{ij}) = 0\,.$$

Plugging (12) into (11) proves case (ii).

We now turn to case (i). Suppose $\{i,j\}\in E$. Eq. (11) implies

$$\begin{aligned}\operatorname{sign}(\theta_{ij})\cdot \mathsf{E}_x(X_{ij}) \ge{}& \operatorname{sign}(\theta_{ij})\cdot \mathsf{E}_x(X_{ij}|D_{ij})\mathsf{P}(D_{ij}) \\ &- |\mathsf{E}_x(X_{ij}|D_{ij}^c)\mathsf{P}(D_{ij}^c)|\,. \end{aligned} \tag{14}$$

The second term has already been bounded in (12). We estimate the first term on the right-hand side of (14):

$$\begin{aligned}&\operatorname{sign}(\theta_{ij})\cdot \mathsf{E}_x(X_{ij}|D_{ij})\mathsf{P}(D_{ij}) \\ &\overset{(a)}{=} 2\operatorname{sign}(\theta_{ij})\big(p^+(x_{\partial i\setminus\{j\}})-p^-(x_{\partial i\setminus\{j\}})\big)\mathsf{P}_x(C_{ij}|D_{ij})\mathsf{P}(D_{ij}) \\ &\overset{(b)}{\ge} 2\big(e^{4|\theta_{ij}|}-1\big)\cdot \tfrac{1}{16}e^{-10d\beta}\mathsf{P}(A_{ij})\mathsf{P}(D_{ij}) \\ &\overset{(c)}{\ge} 2\cdot 4|\theta_{ij}|\tfrac{1}{16}e^{-10d\beta}qe^{-Ld}\,. \end{aligned}$$

Here (a) uses (10), (b) is by (6) and because the reasoning from (8) applies also conditioned on D_{ij} and using the fact that A_{ij} and D_{ij} are independent, and (c) follows

from the inequality $e^x \geq 1+x$, the definition $q = \mathsf{P}(A_{ij})$, and (9). This proves part (i). ∎

We will use the following Bernstein-type submartingale concentration inequality, which can be found for example as an implication of Theorem 27 in [45].

Lemma 4: Let $Z_1, \ldots, Z_n$ be a submartingale adapted to the filtration $(\mathcal{F}_k)_{k\geq 0}$ with $|Z_k - Z_{k-1}| \leq c$ almost surely and $\mathrm{Var}(Z_k|\mathcal{F}_{k-1}) \leq \sigma^2$. Then for all $N \geq 0$ and real t,

$$\mathsf{P}(Z_N - Z_0 \leq -t) \leq \exp\left(-\frac{t^2}{2N\sigma^2 + ct/3}\right).$$

We now prove Theorem 2.

Proof: Recall that $q = [(1-e^{-L/3})e^{-L/3})]^3$. Suppose that $\{i,j\} \in E$. Let ρ denote the lower bound quantity in case (i) of Lemma 3. The inequality $e^{-t} \geq 1-t$ implies that

$$\begin{aligned}\rho &= 2q\Big(|\theta_{ij}| \cdot \tfrac{1}{4} e^{-10d\beta} e^{-Ld} - Ld\Big) \\ &\geq 2q\Big(\frac{\alpha}{4} e^{-10d\beta} e^{-Ld} - Ld\Big) \\ &\geq 2q(4Lde^{-Ld} - Ld) \geq 4qLd\,.\end{aligned}$$

Here we used the bound $L \leq \beta e^{-10\beta}/16d \leq 1/160d$ so $e^{-Ld} \geq e^{-1/160} \geq 3/4$.

The sequence $Z_k = \sum_{\ell=1}^k \mathrm{sign}(\theta_{ij}) X_{ij}^{(\ell)} - k\rho$, $k \geq 1$, is a submartingale adapted to the filtration $(\mathcal{F}_k)_{k\geq 1} = (\sigma^{[0,kL)})_{k\geq 1}$, since by Lemma 3,

$$\mathsf{E}(\mathrm{sign}(\theta_{ij}) X_{ij}^{(k)} | \sigma^{((k-1)L)}) \geq \min_x \{\mathrm{sign}(\theta_{ij}) \mathsf{E}_x X_{ij}\} \geq \rho\,.$$

Let $\overline{X}_{ij} = \frac{1}{k_{\max}} \sum_{k=1}^{k_{\max}} X_{ij}^{(k)}$. Define σ^2 as in Lemma 4, and note that $|Z_k - Z_{k-1}| \leq 2 + \rho \leq 3$. Recalling the choice $\tau = 3Ldq$, by Lemma 4

$$\begin{aligned}\mathsf{P}(\mathrm{sign}(\theta_{ij})\overline{X}_{ij} < \tau) &= \mathsf{P}(Z_{k_{\max}} < k_{\max}(\tau - \rho)) \\ &\leq \mathsf{P}(Z_{k_{\max}} < -k_{\max} Ldq) \\ &\leq \exp\left(-\frac{k_{\max}(Ldq)^2}{2\sigma^2 + Ldq}\right).\end{aligned}$$

It remains to bound $\mathrm{Var}(Z_k|\mathcal{F}_{k-1})$. For this, we observe that $\mathrm{Var}(Z_k|\mathcal{F}_{k-1}) \leq 4 \cdot \mathsf{P}(X_{ij}^{(k)} \neq 0)$ (since $X_{ij}^{(k)} \in \{-2, 0, 2\}$). Now, $\mathsf{P}(X_{ij}^{(k)} \neq 0) \leq \mathsf{P}(C_{ij}) \leq q$, so $\sigma^2 \leq 4q$. We therefore obtain

$$\mathsf{P}(\mathrm{sign}(\theta_{ij})\overline{X}_{ij} < \tau) \leq \exp(-k_{\max} L^2 d^2 q/9)\,,$$

where we used the crude bound $Ld \leq 1$.

If $k_{\max} = 27(L^2d^2q)^{-1}\log p$ then we can take a union bound over the at most $pd/2 \leq p^2$ edges to see that with probability at least $1 - \frac{1}{p}$ we have $E \subseteq \widehat{E}$. We can translate this value for $k_{\max}$ to the time T stated in the theorem by taking T larger than

$$\frac{64^2 \cdot 27 \cdot 6}{\alpha^2} e^{20d\beta} \log p = \frac{27 \cdot 6}{L^2 d^2} \log p \geq \frac{27}{Ld^2 q} \log p = Lk_{\max}\,.$$

The inequality used the estimate which holds for $L \leq 1/2$: $1/q \leq 3e^L L^{-1} \leq 6L^{-1}$.

Next, suppose that $\{i,j\} \notin E$. Lemma 3 states that $|\mathsf{E}_x X_{ij}| \leq 2q(1 - e^{-Ld}) \leq 2Ldq := \rho'$. As before this implies that $Z_k = \sum_{\ell=1}^k X_{ij}^{(\ell)} - k\rho'$, $k \geq 1$, is a supermartingale and $\widetilde{Z}_k = \sum_{\ell=1}^k X_{ij}^{(\ell)} + k\rho'$, $k \geq 1$, is a submartingale. Lemma 4 gives

$$\begin{aligned}\mathsf{P}(|\overline{X}_{ij}| \geq \tau) &\leq \mathsf{P}(Z_{k_{\max}} \geq k_{\max}(\tau - \rho')) \\ &\quad + \mathsf{P}(\widetilde{Z}_{k_{\max}} \leq k_{\max}(\rho' - \tau)) \\ &\leq 2\exp\left(-\frac{k_{\max}(Ldq)^2}{2\sigma^2 + Ldq}\right).\end{aligned}$$

The same bound on σ^2 applies as before, and a union bound over at most $\binom{p}{2}$ non-edges shows that the same $k_{\max}$ (and hence T) specified earlier suffices in order that $\widehat{E} \subseteq E$ with probability $1 - \frac{1}{p}$. ∎

IV. A lower bound on the observation time

Our lower bound derivation is a modification of the proof of Santhanam and Wainwright [46] for the i.i.d. setting. Their construction was based on cliques of size $d+1$, with a single edge removed. When the interaction is ferromagnetic (i.e., $\theta_{ij} \geq 0$), at low temperatures (α, β large enough) the removal of a single edge is difficult to detect and leads to a lower bound.

We use a similar (but not identical) family of models as in [46] to lower bound the observation time required. Start with a graph G_0 consisting of $\lfloor p/(d+1) \rfloor$ cliques of size $d+1$. Suppose that d is odd, and fix a perfect matching on each of the cliques (each matching has cardinality $(d+1)/2$). The vector of parameters θ^0 corresponding to G_0 is obtained by setting $\theta_{ij}^0 = \alpha$ for edges in the matchings, and $\theta_{ij}^0 = \beta$ for edges not in the matchings.

Now for each $\{u,v\}$ in a matching (where $\theta_{uv}^0 = \alpha$) we form the graph G_{uv} by removing the edge $\{u,v\}$ from G_0. There are

$$M = \left\lfloor \frac{p}{d+1} \right\rfloor \left(\frac{d+1}{2}\right) \geq \frac{p}{4}$$

graphs G_{uv} with one edge removed.

This construction is a refinement of the one in [46]: their construction had all edge parameters equal to a single value β, and therefore did not fully capture the effect of some edges being dramatically weaker.

Theorem 5 (Sample complexity lower bound): Suppose the minimax risk is $R_{p,d,n} \leq 1/2$. Then $T = n/p$ satisfies

$$T \geq \frac{e^{2\beta d/3}}{32e^6\alpha} \log p\,.$$

In the remainder of this section we prove Theorem 5. We use the following version of Fano's inequality, which can be found, for example, as Corollary 2.6 in [47]. It gives a lower bound on the error probability (minimax risk in our case) in terms of the KL-divergence between pairs of points in the parameter space, where

KL-divergence between two distributions P and Q on a space $\mathcal{X}$ is defined as

$$D(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}.$$

Lemma 6 (Fano's inequality): Assume that $M \geq 2$ and that Θ contains elements $\theta_0, \theta_1, \ldots, \theta_M$. Let Q_{θ_j} denote the probability law of the observation X under model θ_j. If

$$\frac{1}{M+1}\sum_{j=1}^{M} D(Q_{\theta_j}\|Q_{\theta_0}) \leq \gamma \log M \tag{15}$$

for $0 < \gamma < 1/8$, then the minimax risk for the zero-one loss is bounded as

$$p_e \geq \frac{\log(M+1) - 1}{\log M} - \gamma.$$

A. Bound on KL divergence

In this section we upper bound the KL divergence between the models parameterized by θ^0 and any θ^{uv} (by symmetry of the construction this is the same for every θ^{uv}). It suffices to consider the projection (i.e., marginal) onto the size $d+1$ clique containing u and v, since the KL divergence between these projections is equal to the entire KL divergence. We therefore abuse notation slightly and write P_{θ^0} and $\mathsf{P}_{\theta^{uv}}$ for the Gibbs distributions after projecting onto the relevant clique. Similarly, using θ as a placeholder for either θ^0 or θ^{uv}, we let Q_θ represent the distribution of the observation X, which now consists of samples $\sigma^{(1)}, \ldots, \sigma^{(n)} \in \{-1,+1\}^{d+1}$ as well as node update indices $I^{(1)}, \ldots, I^{(n)} \in [p]$. (We only project the Gibbs measure to the clique, keeping node update indices over the entire original graph.) The initial configuration $\sigma^{(1)}$ is drawn according to the stationary measure P_θ for each model Q_θ. Concretely, with θ representing either θ^0 or θ^{uv},

$$\begin{aligned} &Q_\theta(\sigma^{(1)}, \ldots, \sigma^{(n)}, I^{(1)}, \ldots, I^{(n)}) \\ &\quad = \frac{1}{p^n}\cdot \mathsf{P}_\theta(\sigma^{(1)}) \prod_{l=2}^{n} \mathsf{P}_\theta(\sigma^{(l)}|\sigma^{(l-1)}, I^{(l)}). \end{aligned} \tag{16}$$

Here the factor $1/p^n$ is due to updated node indices being uniformly random at each step. Implicit in the notation is the understanding that $\mathsf{P}_\theta(\sigma^{(l)}|\sigma^{(l-1)}, I^{(l)}) = 0$ if $\sigma^{(l-1)}$ and $\sigma^{(l)}$ differ in any spin other than $I^{(l)}$.

We have the following bound for each of the KL divergence terms in (15) (from which Theorem 5 follows immediately from Lemma 6).

Lemma 7: For each model $Q_{\theta^{uv}}$ on graph G_{uv},

$$D(Q_{\theta^{uv}}\|Q_{\theta^0}) \leq 4\alpha + \frac{n}{p} 18\alpha d e^d e^{-2\beta d/3}.$$

Proof: Using (16) we write

$$\begin{aligned} D(Q_{\theta^{uv}}\|Q_{\theta^0}) &= \mathsf{E}_{X\sim Q_{\theta^{uv}}} \log\frac{Q_{\theta^{uv}}(X)}{Q_{\theta^0}(X)} \\ &:= C_1 + \sum_{l=2}^{n} C_l, \end{aligned} \tag{17}$$

where

$$C_1 = \mathsf{E}_{\sigma\sim\mathsf{P}_{\theta^{uv}}} \log\frac{\mathsf{P}_{\theta^{uv}}(\sigma)}{\mathsf{P}_{\theta^0}(\sigma)} \tag{18}$$

and for $l \geq 2$

$$C_l = \mathsf{E}_{\sigma^{(l)},\sigma^{(l-1)}\sim Q_{\theta^{uv}}} \log\frac{\mathsf{P}_{\theta^{uv}}(\sigma^{(l)}|\sigma^{(l-1)}, I^{(l)})}{\mathsf{P}_{\theta^0}(\sigma^{(l)}|\sigma^{(l-1)}, I^{(l)})}. \tag{19}$$

Note that from any configuration $\sigma^{(l-1)}$, an update to node k other than u or v has ratio of conditional probabilities equal to one (since the neighborhood of k is the same under both models), so each term in (19) is nonzero only if one of the nodes u or v is updated. This introduces a factor $2/p$ for the probability of selecting u or v to update, and by symmetry of the construction we can condition on u updating. Thus,

$$\begin{aligned} &C_l \\ &= \frac{2}{p}\mathsf{E}_{\sigma^{(l)}\sigma^{(l-1)}\sim Q_{\theta^{uv}}} \left[\log\frac{\mathsf{P}_{\theta^{uv}}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)})}{\mathsf{P}_{\theta^0}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)})}\middle| I^{(l)} = u\right]. \end{aligned} \tag{20}$$

When updating node u we have by (2)

$$\begin{aligned} &\frac{\mathsf{P}_{\theta^{uv}}(\sigma_u^{(l)} = +1|\sigma^{(l-1)}, I^{(l)} = u)}{\mathsf{P}_{\theta^0}(\sigma_u^{(l)} = +1|\sigma^{(l-1)}, I^{(l)} = u)} \\ &= \frac{1+\exp(-2\alpha\sigma_v^{(l-1)} - 2\beta\sum_{j\notin\{u,v\}}\sigma_j^{(l-1)})}{1+\exp(-2\beta\sum_{j\notin\{u,v\}}\sigma_j^{(l-1)})} \qquad (21) \\ &= \frac{\exp(2\beta\sum_{j\notin\{u,v\}}\sigma_j^{(l-1)}) + \exp(-2\alpha\sigma_v^{(l-1)})}{\exp(2\beta\sum_{j\notin\{u,v\}}\sigma_j^{(l-1)}) + 1} \qquad (22) \\ &\leq e^{2\alpha}. \qquad (23) \end{aligned}$$

The summations indexed by $j \notin \{u,v\}$ are over nodes in the size $d+1$ clique under consideration. The last inequality follows by observing that the largest value is achieved in (22) when $\sigma_v^{(l-1)} = -1$ and $\sum_{j\notin\{u,v\}}\sigma_j^{(l-1)} \to -\infty$. By symmetry the same bound holds for the ratio of conditional probabilities of $\sigma_u = -1$.

Equation (23) shows that the log-likelihood ratio is always at most 2α. However, it is *typically* roughly $e^{-cd\beta}$, where $c > 0$ is a constant, because the effective magnetic field $\beta\sum_{j\notin\{u,v\}}\sigma_j$ typically has magnitude on the order βd, as shown in Lemma 8 later in this section. Consider the event $U_l = \{\sum\sigma_i^{(l-1)} \geq d/3+2\}$. Applying the inequality $e^{2z} - 1 \leq 7z$ for $0 \leq z \leq 1$ to (22) gives

$$\begin{aligned} &\frac{\mathsf{P}_{\theta^{uv}}(\sigma_u^{(l)} = +1|U_l, I^{(l)} = u)}{\mathsf{P}_{\theta^0}(\sigma_u^{(l)} = +1|U_l, I^{(l)} = u)} \\ &\leq 1 + \frac{e^{2\alpha}-1}{1+\exp(2\beta d/3)} \\ &\leq 1 + 7\alpha\exp e^{-2\beta d/3}, \end{aligned} \tag{24}$$

and also from (2)

$$\mathsf{P}_{\theta^{uv}}(\sigma_u^{(l)} = -1|U_l, I^{(l)} = u) \leq e^{-2\beta d/3}. \tag{25}$$

We now bound each term C_l in (20). Let $\overline{U}_l$ denote the event that $\left|\sum_i \sigma^{(l-1)}\right| \geq d/3+2$. Conditioning on $\overline{U}_l$ gives

$$\begin{aligned}
C_l &= \mathsf{E}_{Q_{\theta^{uv}}} \log \frac{\mathsf{P}_{\theta^{uv}}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)}{\mathsf{P}_{\theta^0}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)} \\
&= \mathsf{E}_{Q_{\theta^{uv}}}\left[\log \frac{\mathsf{P}_{\theta^{uv}}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)}{\mathsf{P}_{\theta^0}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)}\Big|\overline{U}_l\right]\mathsf{P}_{\theta^{uv}}(\overline{U}_l) \\
&\quad + \mathsf{E}_{Q_{\theta^{uv}}}\left[\log \frac{\mathsf{P}_{\theta^{uv}}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)}{\mathsf{P}_{\theta^0}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)}\Big|\overline{U}_l^c\right]\mathsf{P}_{\theta^{uv}}(\overline{U}_l^c) \\
&= \mathsf{E}_{Q_{\theta^{uv}}}\left[\log \frac{\mathsf{P}_{\theta^{uv}}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)}{\mathsf{P}_{\theta^0}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)}\Big|U_l\right]\mathsf{P}_{\theta^{uv}}(\overline{U}_l) \\
&\quad + \mathsf{E}_{Q_{\theta^{uv}}}\left[\log \frac{\mathsf{P}_{\theta^{uv}}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)}{\mathsf{P}_{\theta^0}(\sigma_u^{(l)}|\sigma^{(l-1)}, I^{(l)}=u)}\Big|\overline{U}_l^c\right]\mathsf{P}_{\theta^{uv}}(\overline{U}_l^c)
\end{aligned} \tag{26}$$

The only change in the last equality is replacing $\overline{U}_l$ by U_l in the first conditional expectation, which is justified by symmetry of the model to flipping all the spins. Using (23),(24), (25), $\log(1+x) \leq x$, and $\mathsf{P}_{\theta^{uv}}(\overline{U}_l) \leq 1$, the first term in (26) is bounded by

$$2\alpha e^{-2\beta d/3} + 7\alpha e^{-2\beta d/3} \leq 9\alpha e^{-2\beta d/3}\,.$$

Using (23) and Lemma 8 below, the second term in (26) is bounded by

$$2\alpha \mathsf{P}_{\theta^{uv}}(\overline{U}_l^c) \leq 2\alpha d(3e)^{\frac{d}{3}+1}\exp(-\beta d(d-3)/3)\,.$$

Combining the last two displays gives

$$\begin{aligned}
C_l &\leq 9\alpha e^{-2\beta d/3} + 2\alpha d(3e)^{\frac{d}{3}+1}\exp(-\beta d(d-3)/3) \\
&\leq 9\alpha d e^d e^{-2\beta d/3}
\end{aligned}$$

and adding this quantity n times and multiplying by the factor $2/p$ in (20), we get

$$D(Q_{\theta^{uv}}\|Q_{\theta^0}) \leq C_1 + \frac{n}{p}18\alpha d e^d e^{-2\beta d/3}\,.$$

Now, to bound C_1, it suffices to bound $\mathsf{P}_{\theta^{uv}}(\sigma)/\mathsf{P}_{\theta^0}(\sigma)$. Let $g_{uv}(\sigma) = Z_{uv}\mathsf{P}_{\theta^{uv}}(\sigma)$ and $g_0(\sigma) = Z_0\mathsf{P}_{\theta^0}(\sigma)$, where $Z_{uv} = \sum_\sigma g_{uv}(\sigma)$ and $Z_0 = \sum_\sigma g_0(\sigma)$ are the partition functions for the two models. An argument similar to (23) shows that $e^{-2\alpha}g_{uv}(\sigma) \leq g_0(\sigma) \leq e^{2\alpha}g_{uv}(\sigma)$ for any σ, hence

$$C_1 \leq \log\frac{\mathsf{P}_{\theta^{uv}}(\sigma)}{\mathsf{P}_{\theta^0}(\sigma)} = \log\frac{Z_0\cdot g_{uv}(\sigma)}{Z_{uv}\cdot g_0(\sigma)} \leq 4\alpha\,.$$

Plugging this quantity into the previous displayed equation completes the proof. ∎

Lemma 8: The magnetization $\sum_i \sigma_i$ satisfies

$$\mathsf{P}_{\theta^{uv}}(|\textstyle\sum_i\sigma_i| \leq d/3+1) \leq d(3e)^{\frac{d}{3}+1}\exp(-\beta d(d-3)/3)\,.$$

Proof: Note that $\mathsf{P}_{\theta^{uv}}$ is the stationary measure for the Glauber dynamics governing $Q_{\theta^{uv}}$, so the marginal distribution of each $\sigma^{(l)}$ in the sample $X \sim Q_{\theta^{uv}}$ is $\mathsf{P}_{\theta^{uv}}$.

We first lower bound $\mathsf{P}_{\theta^{uv}}(|\sum_i\sigma_i| > d/3+1)$ by the probability of the all $+1$ or all -1 configuration,

$$\begin{aligned}
&\mathsf{P}_{\theta^{uv}}\left(\Big|\sum_i \sigma_i\Big| > \frac{d}{3}+1\right) \\
&\geq \frac{2}{Z}\exp\left(\beta\frac{(d-1)(d+1)}{2} + \alpha\frac{d-1}{2}\right) \\
&\geq \frac{2}{Z}\exp\left(\beta\cdot\frac{d^2-1}{2}\right).
\end{aligned}$$

Next, by supposing all edges in the clique have parameter β we get the upper bound

$$\begin{aligned}
\mathsf{P}_{\theta^{uv}}\left(\Big|\sum_i\sigma_i\Big| \leq \frac{d}{3}+1\right) &\leq \frac{2d}{Z}\binom{d}{\frac{d}{3}+1}\exp(\beta d^2/9 + \beta d/2) \\
&\leq (3e)^{\frac{d}{3}+1}\frac{2d}{Z}\exp(\beta d^2/9+\beta d/2)\,,
\end{aligned}$$

where the second inequality follows from $\binom{n}{k} \leq \left(\frac{n\cdot e}{k}\right)^k$ and $(3e)^{d/3} \leq e^d$. Taking the ratio of the last two displayed quantities gives the desired inequality. ∎

V. Discussion

The main message of this paper is that observing dynamics over time is quite natural in many settings, and that access to such observations leads to a simple algorithm for estimating the graph underlying an Ising model. We expect that similar results can be derived (with suitable modifications) for samples generated from local Markov chains other than the Glauber dynamics, and for non-binary pairwise graphical models. Several other generalizations are plausible; for instance, it would be interesting to consider the situation where one only observes samples intermittently.

Acknowledgments

We are grateful to Mina Karzand, Kuang Xu, and Luis Voloch for helpful comments on a draft of the paper, and to Vijay Subramanian for an interesting discussion on coordination games. This work was supported in part by NSF grants CMMI-1335155 and CNS-1161964, and by Army Research Office MURI Award W911NF-11-1-0036.

References

[1] G. Bresler, D. Gamarnik, and D. Shah, "Hardness of parameter estimation in graphical models," in *NIPS*, 2014.

[2] A. Montanari, "Computational Implications of Reducing Data to Sufficient Statistics," *ArXiv e-prints*, Sept. 2014.

[3] J. Bento and A. Montanari, "Which graphical models are difficult to learn?," in *NIPS*, 2009.

[4] A. Sly and N. Sun, "The computational hardness of counting in two-spin models on d-regular graphs," in *FOCS*, 2012.

[5] P. Abbeel, D. Koller, and A. Y. Ng, "Learning factor graphs in polynomial time and sample complexity," *JMLR*, 2006.

[6] G. Bresler, E. Mossel, and A. Sly, "Reconstruction of Markov random fields from samples: Some observations and algorithms," in *RANDOM*, 2008.

[7] I. Csiszár and Z. Talata, "Consistent estimation of the basic neighborhood of markov random fields," *The Annals of Statistics*, pp. 123–145, 2006.

[8] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. on Info. Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[9] S. Dasgupta, "Learning polytrees," in *UAI*, 1999.

[10] N. Srebro, "Maximum likelihood bounded tree-width Markov networks," in *UAI*, 2001.

[11] A. Anandkumar, F. Huang, D. J. Hsu, and S. M. Kakade, "Learning mixtures of tree graphical models," in *NIPS*, 2012.

[12] G. Bresler, D. Gamarnik, and D. Shah, "Structure learning of antiferromagnetic ising models," in *NIPS*, 2014.

[13] D. Gamarnik, "Correlation decay method for decision, optimization, and inference in large-scale networks," *Tutorials in Operations Research, INFORMS*, 2013.

[14] R. L. Dobrushin, "Prescribing a system of random variables by conditional distributions," *Theory of Probability & Its Applications*, vol. 15, no. 3, pp. 458–486, 1970.

[15] R. L. Dobrushin and S. B. Shlosman, "Constructive criterion for the uniqueness of Gibbs field," in *Statistical physics and dynamical systems*, pp. 347–370, Springer, 1985.

[16] J. Salas and A. D. Sokal, "Absence of phase transition for antiferromagnetic Potts models via the dobrushin uniqueness theorem," *Journal of Statistical Physics*, vol. 86, no. 3-4, pp. 551–579, 1997.

[17] D. Gamarnik, D. A. Goldberg, and T. Weber, "Correlation decay in random decision networks," *Mathematics of Operations Research*, vol. 39, no. 2, pp. 229–261, 2013.

[18] D. Gamarnik and D. Katz, "Correlation decay and deterministic fptas for counting list-colorings of a graph," in *SODA*, 2007.

[19] A. Bandyopadhyay and D. Gamarnik, "Counting without sampling: Asymptotics of the log-partition function for certain statistical physics models," *Random Structures & Algorithms*, vol. 33, no. 4, pp. 452–479, 2008.

[20] D. Weitz, "Counting independent sets up to the tree threshold," in *STOC*, 2006.

[21] P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai, "Greedy learning of Markov network structure," in *Allerton*, 2010.

[22] A. Ray, S. Sanghavi, and S. Shakkottai, "Greedy learning of graphical models with small girth," in *Allerton*, 2012.

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

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

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

[26] S.-I. Lee, V. Ganapathi, and D. Koller, "Efficient structure learning of Markov networks using ℓ_1-regularization," in *NIPS*, pp. 817–824, 2006.

[27] A. Jalali, C. Johnson, and P. Ravikumar, "On learning discrete graphical models using greedy methods," *NIPS*, 2011.

[28] A. Jalali, P. Ravikumar, V. Vasuki, and S. Sanghavi, "On learning discrete graphical models using group-sparse regularization," in *AISTATS*, 2011.

[29] D. W. Stroock and B. Zegarlinski, "The logarithmic Sobolev inequality for discrete spin systems on a lattice," *Comm. in Mathematical Physics*, vol. 149, no. 1, pp. 175–193, 1992.

[30] M. Dyer, A. Sinclair, E. Vigoda, and D. Weitz, "Mixing in time and space for lattice spin systems: A combinatorial view," *Random Structures & Algorithms*, vol. 24, no. 4, pp. 461–479, 2004.

[31] F. Martinelli and E. Olivieri, "Approach to equilibrium of Glauber dynamics in the one phase region," *Comm. in Mathematical Physics*, vol. 161, no. 3, pp. 447–486, 1994.

[32] P. Netrapalli and S. Sanghavi, "Learning the graph of epidemic cascades," in *SIGMETRICS*, 2012.

[33] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," *arXiv preprint arXiv:1105.0697*, 2011.

[34] S. Myers and J. Leskovec, "On the convexity of latent social network inference," in *NIPS*, 2010.

[35] A. Anandkumar, A. Hassidim, and J. Kelner, "Topology discovery of sparse random graphs with few participants," *SIGMETRICS*, 2011.

[36] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?," *IEEE Trans. on Info. Theory*, vol. 57, no. 8, pp. 5163–5181, 2011.

[37] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Efficiently spotting the starting points of an epidemic in a large graph," *Knowledge and information systems*, 2014.

[38] D. Aldous and U. Vazirani, "A Markovian extension of valiant's learning model," in *FOCS*, 1990.

[39] P. L. Bartlett, P. Fischer, and K.-U. Höffgen, "Exploiting random walks for learning," in *COLT*, 1994.

[40] N. H. Bshouty, E. Mossel, R. O'Donnell, and R. A. Servedio, "Learning dnf from random walks," *Journal of Computer and System Sciences*, vol. 71, no. 3, pp. 250–265, 2005.

[41] D. Gamarnik, "Extension of the PAC framework to finite and countable Markov chains," *IEEE Trans. on Info. Theory*, vol. 49, no. 1, pp. 338–345, 2003.

[42] A. Montanari and A. Saberi, "The spread of innovations in social networks," *Proceedings of the National Academy of Sciences*, vol. 107, no. 47, pp. 20196–20201, 2010.

[43] R. Berry and V. G. Subramanian, "Spotting trendsetters: Inference for network games," in *Allerton*, 2012.

[44] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani, "Random generation of combinatorial structures from a uniform distribution," *Theoretical Computer Science*, vol. 43, pp. 169–188, 1986.

[45] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: a survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, 2006.

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

[47] A. B. Tsybakov, *Introduction to nonparametric estimation*, vol. 11. Springer, 2009.