Efficiently Learning Ising Models on Arbitrary Graphs

[Extended Abstract]

Guy Bresler
Laboratory for Information and Decision Systems
Department of EECS
Massachusetts Institute of Technology
gbresler@mit.edu

ABSTRACT

We consider the problem of reconstructing the graph underlying an Ising model from i.i.d. samples. Over the last fifteen years this problem has been of significant interest in the statistics, machine learning, and statistical physics communities, and much of the effort has been directed towards finding algorithms with low computational cost for various restricted classes of models. Nevertheless, for learning Ising models on general graphs with p nodes of degree at most d, it is not known whether or not it is possible to improve upon the p^d computation needed to exhaustively search over all possible neighborhoods for each node.

In this paper we show that a simple greedy procedure allows to learn the structure of an Ising model on an arbitrary bounded-degree graph in time on the order of p^2. We make no assumptions on the parameters except what is necessary for identifiability of the model, and in particular the results hold at low-temperatures as well as for highly non-uniform models. The proof rests on a new structural property of Ising models: we show that for any node there exists at least one neighbor with which it has a high mutual information.

Categories and Subject Descriptors

G.3 [**Mathematics of computing**]: Probability and statistics—*Distribution functions*; F.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

Keywords

Ising model; structure learning; Markov random field

1. INTRODUCTION

Undirected graphical models, or Markov random fields, are a general and powerful framework for reasoning about high dimensional distributions and are at the core of modern statistical inference. The joint probability distribution specified by such a model factorizes according to an underlying graph, and the absence of edges encodes conditional independence [40]. The graph structure captures the computational aspect inherent in tasks of statistical inference including computing marginals, maximum *a posteriori* assignments, the partition function, or sampling from the distribution. In addition to their statistical relevance, such computations on graphical models include as special cases many combinatorial optimization and counting problems.

The Ising model is a Markov random field having binary variables and pairwise potential functions. The Ising model has a long and celebrated history starting with its introduction by statistical physicists as a model for spin systems in order to understand the phenomena of *phase transition* [33, 12]. It has since been used across a wide spectrum of application domains including finance, social networks, computer vision, biology, and signal processing. The understanding of the computational tractability of inference (computing the partition function and sampling) has recently seen significant progress [36, 74, 63, 64, 65, 30].

The inverse problem of learning models from data is equally important. Once the underlying graph is known it is relatively easy to estimate the parameters, hence the focus is largely on the task of *structure learning*, i.e., estimating the graph. Study of this problem was initiated by Chow and Liu in their seminal 1968 paper [13], which gave a greedy algorithm for learning *tree-structured* Markov random fields with runtime on the order of p^2 for graphs on p nodes. They showed that the maximum likelihood graph is a maximum-weight spanning tree, where each edge has weight equal to the mutual information of the variables at its endpoints. The maximum-likelihood tree can thus be found by a greedy algorithm, for example using Kruskal's or Prim's algorithms, and the running-time of the Chow-Liu algorithm is dominated by the time required to compute the mutual information between all pairs of nodes in the graph.

For graphs with loops the problem is much more challenging for two reasons: a node and its neighbor can be *marginally independent* due to indirect path effects, and moreover, this difficulty is compounded by presence of long-range correlations in the model, in which case distant nodes can be more correlated than nearby nodes. As discussed next in Subsection 1.1, a basic first-order question has remained unanswered: it is not known if it is possible to learn the structure of Ising models on general graphs with p nodes of degree at most d in time less than p^d. This is roughly the time required to exhaustively search over all $\binom{p}{d}$ possible neighborhoods of a node and for each such candidate

STOC'15, June 14–17, 2015, Portland, Oregon, USA.

ACM 978-1-4503-3536-2/15/06 ...$15.00.
http://dx.doi.org/10.1145/2746539.2746631.

neighborhood test whether or not the implied conditional independence holds [11].

In this paper we show that despite these challenges, a simple greedy algorithm can learn arbitrary Ising models on p nodes of maximum degree d in time $\widetilde{\mathcal{O}}(p^2)$, where the $\widetilde{\mathcal{O}}(\cdot)$ notation hides a factor $\log p$ as well as a constant depending (doubly-exponentially) on d. Exponential dependence on d is unavoidable as the number of samples must itself increase exponentially in d [60], but doubly-exponential dependence on d is probably suboptimal. The implication of our result is that learning Ising models on *arbitrary graphs* of bounded degree has essentially the same computational complexity as learning such models on *tree graphs*. The proof rests on a new basic property of Ising models: we show that for any node, there exists at least one neighbor with which it has high mutual information, even conditional on any subset of nodes.

1.1 Complexity of graphical model learning

A number of papers, including [1], [11], and [17], have suggested to find 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 work 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. We will think of efficient algorithms as having an exponent $c(d)$ that is bounded by a constant independent of d[1].

Previous works proposing efficient algorithms either restrict the graph structure or the nature of the interactions between variables. Chow and Liu [13] made a model restriction of the first type, assuming that the graph is a tree; generalizations include to polytrees [19], hypertrees [66], tree mixtures [3], and others. Among the many possible assumptions of the second type, the correlation decay property (CDP) is distinguished: until the recent paper [10], all existing efficient algorithms required the CDP [8] or had conditions that were not interpretable in terms of model parameters. Informally, a graphical model is said to have the CDP if any two variables σ_i and σ_j are asymptotically independent as the graph distance between i and j increases. The CDP is known to hold for a number of graphical models in the so-called high-temperature regime, including Ising, hard-core lattice gas, Potts (multinomial), and others (see the survey article [28] as well as, e.g., [22, 23, 59, 74, 29, 6]).

It was first observed in [11] that it is possible to efficiently learn (in time $\widetilde{\mathcal{O}}(p^2)$) models with (exponential) decay of correlations, under the additional assumption that neighboring variables have correlation bounded away from zero (as is true for the ferromagnetic Ising model in the high temperature regime). A variety of other papers including [51, 56, 4] give alternative algorithms, but also require the CDP to guarantee efficiency. Most structure learning algorithms that do not explicitly require the CDP are based on convex optimization, such as Ravikumar, Wainwright, and Lafferty's [54] approach using ℓ_1-regularized node-wise logistic regression. This algorithm has complexity $\mathcal{O}(p^4)$; while it is shown to work under certain incoherence conditions that seem distinct from the CDP, Bento and Montanari [8] established through a careful analysis that the algorithm fails for simple families of ferromagnetic Ising models without the CDP. Other convex optimization-based algorithms (e.g., [41, 34, 35]) assume similar incoherence conditions that are difficult to interpret in terms of model parameters, and likely also require the CDP.

It is noteworthy that most computationally efficient *sampling* algorithms (which happen to be based on MCMC) require a notion of temporal mixing, and this is closely related to spatial mixing or a version of the CDP (see, e.g., [67, 24, 43, 74]). Thus, under a class of *mixing conditions*, we can both generate samples efficiently as well as learn graphical models efficiently from i.i.d. samples. For antiferromagnetic Ising models on general bounded degree graphs, one has the striking converse statement that generating samples or approximating the partition function becomes intractable (NP-hard) precisely at the point where the CDP no longer holds [65].

Because all known efficient algorithms required the CDP, and because the Ising model exhibits dramatically different macroscopic behavior with versus without the CDP (and this determines computational tractability of sampling), Bento and Montanari [8] posed the question of whether or not the CDP is necessary for tractable structure learning. A partial answer was given in [10], by demonstrating that a family of antiferromagnetic Ising models on general graphs can be learned efficiently despite strongly violating the CDP. Thus any relationship between the complexity of sampling (or computing the partition function) and the problem of *structure learning* from i.i.d. samples seems tenuous, and this is corroborated by the results of this paper. In contrast, the recent papers [9] and [47] demonstrate an algorithmic connection between *parameter estimation* from *minimal sufficient statistics* for a model on a known graph and computation of the partition function.

1.2 Results

We prove that the graph structure of an arbitrary Ising model on a bounded degree graph can be learned efficiently from i.i.d. samples. Before discussing the algorithm, we first state the main conceptual contribution of the paper, namely identifying a new structural property of Ising models. Given the structural result the algorithm is almost obvious, and indeed it can be interpreted as a generalization of Chow and Liu's 1968 greedy algorithm for learning trees to models on arbitrary graphs.

Proposition 1.1 (Structural property – informal)**.** *Let G be a graph on p nodes of maximum degree d, and consider an Ising model on G. Then for any node $u \in \mathcal{V}$ there exists a neighbor i such that the mutual information between i and u is at least some constant that is independent of p.*

The property as stated is actually a consequence of Proposition 5.3: we allow to condition on an arbitrary set of nodes, and instead of mutual information, we use a certain conditional influence measure. As shown in Section 5, this influence provides a lower bound on the mutual information.

The proof of the structural property starts with the fact that at any finite temperature an Ising model on a bounded degree graph has nontrivial randomness in each variable. Note that this local property holds regardless of whether or

[1]This notion of efficiency is known as fixed-parameter tractability [52].

not the model satisfies the correlation decay property (which is a global property). The randomness in the neighbors of a given node and an anti-concentration argument show that the neighbors together influence the node's value in a nontrivial way, and this implies that at least one neighbor has nontrivial influence. We remark that anti-concentration has played a key role in a variety of recent structural results (e.g., [21, 31]).

Our algorithm guarantee is stated in the following theorem.

Theorem 1.2 (Algorithm performance – informal)**.** *Consider an Ising model on an arbitrary graph on p nodes of maximum degree at most d. Given $n = f(d)\log p$ samples from the model, where the constant $f(d)$ is a function of the range of interaction strengths and d, it is possible to learn the underlying graph in time $f(d)p^2 \log p$.*

The factor $f(d)$ in the sample complexity and runtime of our algorithm depends doubly-exponentially on d, in contrast to the necessary exponential dependence discussed in Subsection 2.2. While the focus of this paper is not on sample complexity, our algorithm does obtain (with a suboptimal constant) the optimal logarithmic dependence on p.

Our algorithm is described in Section 4 and a more detailed statement of the theorem (with full dependence on constants) is given as Theorem 4.1. The algorithm is extremely simple: in order to find the neighborhood of a node u, it greedily (according to a measure of conditional influence) adds nodes one-by-one to form a constant-size superset of the neighborhood (pseudo-neighborhood). The idea of adding spurious nodes to form a superset is not new, but stands in contrast to algorithms that attempt to add only correct nodes.

Adding non-neighbors to the pseudo-neighborhood is beneficial: in models with long-range correlations a non-neighbor i with high influence on u (or high mutual information) contains a lot of information about many *other non-neighbors*, so conditioning on i effectively eliminates many non-neighbors from consideration. This allows us to use a potential argument, whereby each added node reduces the conditional entropy of the node u by some constant, and this bounds the size of the pseudo-neighborhood. The pseudo-neighborhood can then be easily pruned to remove non-neighbors.

We next mention a few connections to other work and then in Section 2 define the Ising model and structure learning problem. Section 3 introduces the notion of influence we use and states a lemma showing that empirical estimates are close to the population averages. Section 4 presents the algorithm with performance guarantee stated in Theorem 4.1. Section 5 contains proofs of correctness and runtime, as well as a statement of our structural result, Proposition 5.3. The proposition is proved in Sections 6 and 7. Finally, Appendix A contains the proof of the lemma from Section 3 and Section 8 discusses possible extensions.

1.3 Other related work

Since the 1980's, Hinton and others have studied the problem of learning Ising models under the name of learning Boltzmann machines [2, 32, 70]. Most approaches for learning Boltzmann machines do not assume a sparse underlying graph and attempt to find parameters directly, using gradient optimization methods. Ising models are used to model neuronal networks and protein interaction networks, and the learning problem is of interest in that context [61, 14, 48, 73]. In the statistical physics community the problem is known as the inverse Ising problem and a variety of interesting non-rigorous methods have been proposed, including based on truncation of expansions relating couplings to mean parameters or entropies [62, 15, 58, 42], message-passing [45, 57, 5], and others [20].

Structure learning of graphical models has been studied in the statistics and machine learning communities as a problem in high-dimensional statistics. Broadly, in high-dimensional statistics one wishes to estimate a high-dimensional object from samples, where the number of samples is far less than the dimensionality of the parameter space. Solving this ill-determined problem requires that there be some underlying structure. In our case the graph underlying the Markov random field is a sparse graph. As discussed in [50], optimization of regularized objective functions has been a popular approach to many problems in high-dimensional statistics including sparse linear regression, low-rank matrix completion, inferring rankings, as well as learning graphical models (both binary and Gaussian) [54, 44, 26, 55]. Such general methodology based on optimizing likelihood or (pseudo-likelihood) has failed thus-far to learn Ising models on general graphs.

Greedy methods have been prominent in the statistics literature for decades and recently many papers have appeared giving theoretical guarantees for sparse linear regression and support recovery problems. Our greedy algorithm is essentially equivalent to forward step-wise regression with a pruning/cleaning step, called "screen and clean" by Roeder and Wasserman in the linear regression setting [72]. Greedy approaches have been studied in the context of approximation theory [69], as an interpretation of boosting [27], and for sparse linear regression [18, 49, 76]. In the graphical model setting, several papers including [34, 56, 51] have analyzed greedy algorithms under various further assumptions on the model.

The theoretical computer science community has made progress on learning a variety of high-dimensional probability distributions from samples, including mixtures of Gaussians [46, 7]. But there is a more intriguing connection to work on learning function classes. In an Ising model, the conditional distribution of a node given its neighbors is specified by a logistic function, which is a soft version of a linear threshold function (LTF). Thus our algorithm effectively learns soft LTFs over a complicated joint distribution. Arguments based on boolean Fourier analysis have played a major role in learning boolean functions over uniformly random examples and also over product distributions [37], but due to the joint dependencies in an Ising model, it is not obvious how to apply Fourier analysis in this setting. Our structural result, nevertheless, is at a high level analogous to the statement that LTFs have non-trivial total degree-one Fourier mass (see, e.g., [53]). Other recent works learn LTFs over non-product distributions, including log-concave [38] and sub-Gaussian or sub-exponential [39]. These assumptions are badly violated by Ising models in the low-temperature regime (with long-range correlations), but the bounded graph degree assumption makes our soft LTFs depend on few variables, and this makes learning tractable.

2. PRELIMINARIES

2.1 Ising model

We consider the Ising model on a graph $G = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = p$. 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 $i \in \mathcal{V}$ is associated a binary random variable (spin) X_i. Each configuration of spins $x \in \{-,+\}^{\mathcal{V}}$ ('$-$' and '$+$' are used as shorthand for $-1, +1$) is assigned probability according to the probability mass function

$$P(x) = \exp\Big(\sum_{\{i,j\}\in\mathcal{E}} \theta_{ij} x_i x_j + \sum_{i\in\mathcal{V}} \theta_i x_i - \Phi(\theta) \Big). \qquad (1)$$

Here $\Phi(\theta)$ is the log-partition function or normalizing constant. The distribution is parameterized by the vector $\theta = \{\theta_{ij}\}_{\{i,j\}\in\mathcal{E}} \cup \{\theta_i\}_{i\in\mathcal{V}} \in \mathbb{R}^{\mathcal{E}\cup\mathcal{V}}$, consisting of edge couplings and node-wise external fields. The edge couplings are assumed to satisfy $\alpha \le |\theta_{ij}| \le \beta$ for all $\{i,j\} \in \mathcal{E}$ for some constants $0 < \alpha \le \beta$ and the external fields are assumed to satisfy $|\theta_i| \le h$ for all $i \in \mathcal{V}$. The bounds α, β, h are necessary for model identifiability, and as shown in [60] and discussed briefly in the next subsection, must appear in the sample complexity.

We can alternatively think of $\theta \in \mathbb{R}^{\binom{p}{2}+p}$, with $\theta_{ij} = 0$ if $\{i,j\} \notin \mathcal{E}$. For a graph G, let

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

be the set of valid parameter vectors corresponding to G.

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. The conditional probability of $X_u = +$ given the states of all the other nodes $\mathcal{V}\setminus\{u\}$ can thus be written as:

$$\mathsf{P}(X_u = + \,|X_{\mathcal{V}\setminus\{u\}} = x_{\mathcal{V}\setminus\{u\}}) = \mathsf{P}(X_u = + \,|X_{\partial u} = x_{\partial u}) = \frac{\exp\left(2\sum_{i\in\partial u}\theta_{ui}x_i + \theta_u\right)}{1+\exp\left(2\sum_{i\in\partial u}\theta_{ui}x_i + \theta_u\right)}. \qquad (2)$$

A useful property of bounded degree models is that the conditional probability of a spin is always bounded away from 0 and 1. The proof of this statement is immediate from (2) by conditioning on the neighbors of u and using the tower property of conditional expectation.

Lemma 2.1 (Conditional randomness)**.** *For any node $u \in \mathcal{V}$, subset $\mathcal{S} \subseteq \mathcal{V}\setminus\{u\}$, and any configuration $x_{\mathcal{S}} \in \{-,+\}^{|\mathcal{S}|}$,*

$$\min\{\mathsf{P}(X_u = +|X_{\mathcal{S}} = x_{\mathcal{S}}), \mathsf{P}(X_u = -|X_{\mathcal{S}} = x_{\mathcal{S}})\} \ge \frac{1}{2}e^{-2(\beta d + h)} := \delta.$$

The quantity δ appears throughout the paper.

2.2 Graphical model learning

Denote the set of all graphs on p nodes of degree at most d by $\mathcal{G}_{p,d}$. For some graph $G \in \mathcal{G}_{p,d}$ and parameters $\theta \in \Omega_{\alpha,\beta,h}(G)$, one observes configurations $X^{(1)}, \dots, X^{(n)} \in \{-,+\}^p$ sampled independently from the Ising model (1). A *structure learning algorithm* is a (possibly randomized) map

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

taking n samples $X^{1:n} = X^{(1)}, \dots, X^{(n)}$ to a graph $\phi(X^{1:n})$. The statistical performance of a structure learning algorithm will be measured using the zero-one loss, meaning that the exact underlying graph must be learned. The risk, or expected loss, under some vector $\theta \in \Omega_{\alpha,\beta,h}(G)$ of parameters corresponding to a graph $G \in \mathcal{G}_{p,d}$ is given by the probability of reconstruction error

$$\mathsf{P}_\theta(\phi(X^{1:n}) \ne G),$$

and for given α, β, h, p, d, the maximum risk is

$$\sup_{\substack{G\in\mathcal{G}_{p,d}\\ \theta\in\Omega_{\alpha,\beta,h}(G)}} \mathsf{P}_\theta(\phi(X^{1:n}) \ne G).$$

Our goal is to find an algorithm with maximum risk (probability of error) tending to zero as $p \to \infty$, using the fewest possible number of samples n. This notion of performance is rather stringent, but also robust, being worst-case over the entire class of graphs $\mathcal{G}_{p,d}$ and parameters $\theta \in \Omega_{\alpha,\beta,h}(G)$. A lower bound on the number of samples necessary was obtained by Santhanam and Wainwright in Theorem 1 of [60]:

$$n \ge \frac{e^{\beta d}\log\left(\frac{pd}{4} - 1\right)}{4\alpha d e^{\alpha}}.$$

This means in particular that exponential dependence of the sample complexity (and hence runtime) on the quantity βd is unavoidable.

3. MEASURING THE INFLUENCE OF A VARIABLE

Our algorithm uses a certain *conditional influence* of a variable on another variable. For nodes $u, i \in \mathcal{V}$, subset of nodes $\mathcal{S} \in \mathcal{V}\setminus\{u,i\}$ and configuration $x_{\mathcal{S}} \in \{-,+\}^{\mathcal{S}}$, define

$$\nu_{u|i;x_{\mathcal{S}}} := \mathsf{P}(X_u = +|X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) - \mathsf{P}(X_u = +|X_i = -, X_{\mathcal{S}} = x_{\mathcal{S}}).$$

We also use a quantity we call the *average conditional influence*, which is obtained by performing a weighted average of $|\nu_{u|i;X_{\mathcal{S}}}|$ over random configurations $X_{\mathcal{S}}$:

$$\nu^{\mathsf{avg}}_{u|i;\mathcal{S}} := \mathsf{E}\big(\lambda_i(X_{\mathcal{S}}) \cdot |\nu_{u|i;X_{\mathcal{S}}}|\big).$$

The weights are given by the function

$$\lambda_i(x_{\mathcal{S}}) = 2 \cdot \mathsf{P}(X_i = +|X_{\mathcal{S}} = x_{\mathcal{S}})\mathsf{P}(X_i = -|X_{\mathcal{S}} = x_{\mathcal{S}}).$$

The average influence $\nu^{\mathsf{avg}}_{u|i;\mathcal{S}}$ is essentially equivalent to the conditional mutual information $I(X_u; X_i|X_{\mathcal{S}})$.

By the Markov property (2), the influence is zero for non-neighbors conditional on neighbors, that is, for any $x_{\mathcal{S}}$,

$$\nu_{u|i;x_{\mathcal{S}}} = 0 \qquad \text{for all } i \in \mathcal{V}\setminus\{u,\mathcal{S}\} \text{ if } \partial u \subseteq \mathcal{S}. \qquad (3)$$

This implies the same statement for $\nu^{\mathsf{avg}}_{u|i;\mathcal{S}}$. Our structural result, Proposition 5.3, shows that $\nu^{\mathsf{avg}}_{u|i;\mathcal{S}}$ is bounded below for at least *one neighbor* $i \in \partial u \setminus \mathcal{S}$ if $\mathcal{S}$ does not already contain the neighborhood ∂u. Thus computing $\nu^{\mathsf{avg}}_{u|i;\mathcal{S}}$ allows to determine if $\partial u \subset \mathcal{S}$ or not, and our algorithm given in Section 4 is based on this idea.

Remark 3.1. Other works using a "conditional independence test", for example [4, 75], use a similar measure of influence amounting to $\min_{x_{\mathcal{S}}} |\nu_{u|i;x_{\mathcal{S}}}|$. We do not take the minimum over configurations $x_{\mathcal{S}}$, as there is no guarantee

that $\min_{x_S} |\nu_{u|i;x_S}|$ is nonzero for any neighbor $i \in \partial u$: each X_i can be marginally independent of X_u conditional on *some* configuration x_S.

The *empirical conditional influence* $\widehat{\nu}_{u|i;x_S}$ replaces the probability measure by the empirical measure:

$$\widehat{\nu}_{u|i;x_S} := \widehat{\mathsf{P}}(X_u = +|X_i = +, X_S = x_S) - \widehat{\mathsf{P}}(X_u = +|X_i = -, X_S = x_S)\,,$$

where for $S, T \subseteq V$,

$$\widehat{\mathsf{P}}(X_T = x_T|X_S = x_S) = \frac{\widehat{\mathsf{P}}(X_T = x_T, X_S = x_S)}{\widehat{\mathsf{P}}(X_S = x_S)} \qquad \text{and}$$

$$\widehat{\mathsf{P}}(X_S = x_S) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{\{X_S^{(t)} = x_S\}}\,.$$

Like before we define an averaged version (with average taken according to the empirical measure):

$$\widehat{\nu}^{\text{avg}}_{u|i;S} := \mathsf{E}_{X_S \sim \widehat{\mathsf{P}}}\Big(\widehat{\lambda}_i(X_S) \cdot \widehat{\nu}_{u|i;X_S}\Big)\,,$$

where

$$\widehat{\lambda}_i(x_S) = 2 \cdot \widehat{\mathsf{P}}(X_i = +|X_S = x_S)\widehat{\mathsf{P}}(X_i = -|X_S = x_S)\,.$$

It will be necessary that these empirical influences are sufficiently accurate. Let $\mathcal{A}(\ell, \epsilon)$ denote the event that empirical influences with conditioning set up to size ℓ are accurate to within an additive ϵ:

$$\mathcal{A}(\ell, \epsilon) = \{|\nu^{\text{avg}}_{u|i;S} - \widehat{\nu}^{\text{avg}}_{u|i;S}| \le \epsilon \text{ for all } \{u, i\} \subset V, S \subset V \setminus \{u, i\}, |S| \le \ell\}\,.$$

Lemma 3.2. *Recall the notation $\delta := \frac{1}{2}e^{-2(\beta d+h)}$ and suppose $\ell \le p/4 - 2$. If the number of samples n is at least $\frac{144(\ell+3)}{\epsilon^2\delta^{2\ell}} \log \frac{p}{\zeta}$, then $\mathsf{P}(\mathcal{A}(\ell, \epsilon)) \ge 1 - \zeta$.*

The proof of the lemma follows from Azuma's inequality and is similar to Theorem 2 of [11]. We give the proof in Appendix A.

4. ALGORITHM

We now describe the structure learning algorithm, which learns the neighborhood of an arbitrary individual node $u \in V$. Algorithm LearnNbhd takes as input the node u whose neighborhood ∂u we wish to learn as well as the data $X^{1:n} = X^{(1)}, \ldots, X^{(n)}$ and a threshold parameter τ. The first step is to construct a *superset* S (which we call a pseudo-neighborhood) of the neighborhood ∂u. This is accomplished by greedily adding to S the node i with highest conditional influence $\widehat{\nu}^{\text{avg}}_{u|i;S}$, until there are no nodes i with $\widehat{\nu}^{\text{avg}}_{u|i;S} \ge \tau$. (To simplify the description we set $\widehat{\nu}^{\text{avg}}_{u|i;S} = 0$ if $i \in S \cup \{u\}$.) At each step the conditional influences are computed with respect to the current set S.

As will become apparent in Subsection 5.1, inclusion of non-neighbors is important, as it allows to use a potential argument to show that the constructed pseudo-neighborhood is not too large. Concretely, by definition of the algorithm and a simple lemma relating influence to mutual information, adding a node to the pseudo-neighborhood S always reduces by at least τ the conditional entropy $H(X_u|X_S)$ of the variable X_u whose neighborhood we are trying to find. The entropy is non-negative and was initially at most one (since X_u is binary), so this bounds the size of S.

The correctness of the algorithm relies on Proposition 5.3 of Subsection 5.2, which shows that there is always at least one neighbor with influence above a constant, and we set τ equal to this constant. This implies that the algorithm does not terminate before all the neighbors are added. Finally, after construction of the pseudo-neighborhood, the algorithm removes those nodes with low average influence. Proposition 5.3 is again used to show that no neighbors are removed.

Algorithm 1 LearnNbhd$(X^{(1)}, \ldots, X^{(n)}, \tau, u)$

Pseudo-neighborhood:
1. Let $S = \varnothing$
2. Let $(i^*, \eta^*) = (\arg\max_i \widehat{\nu}^{\text{avg}}_{u|i;S}, \max_i \widehat{\nu}^{\text{avg}}_{u|i;S})$
3. If $\eta^* \ge \tau$, then add i^* to S
4. Else goto Step 6
5. Repeat Steps 2 to 4

Pruning:
6. For each $i \in S$: if $\widehat{\nu}^{\text{avg}}_{u|i;S\setminus\{i\}} < \tau$, then remove i
7. Output S

Theorem 4.1. *Let $G \in \mathcal{G}_{p,d}$ and $\theta \in \Omega_{\alpha,\beta,h}(G)$. Let $\delta = \frac{1}{2}e^{-2(\beta d+h)}$ and define*

$$\tau^* = \frac{\alpha^2\delta^{4d+1}}{16d\beta}\,, \qquad \epsilon^* = \frac{\tau^*}{2}\,, \qquad \ell^* = \frac{2}{(\tau^* - \epsilon^*)^2} = \frac{8}{(\tau^*)^2}\,.$$

Suppose we observe n samples $X^{(1)}, \ldots, X^{(n)}$, for

$$n \ge \frac{144(\ell^* + 3)}{(\epsilon^*)^2\delta^{2\ell^*}} \log \frac{p}{\zeta} = \exp\{c\alpha^{-c'}e^{c''d(\beta d+h)}\} \cdot \log \frac{p}{\zeta}\,,$$

where c, c', c'' denote numerical constants. Then with probability at least $1 - \zeta$, the structure learning algorithm LearnNbhd$(X^{1:n}, \tau^*, u)$ *returns the correct neighborhood ∂u for all $u \in V$ and the runtime for each of the p nodes is (for a numerical constant C)*

$$C\ell^* pn = \mathcal{O}(p \log p)\,.$$

Remark 4.2. As stated, the algorithm has probability $1-\zeta$ of both returning the correct neighborhoods for all nodes and having the claimed runtime. Obviously, the algorithm can be terminated if the runtime exceeds the stated value, giving a deterministic guarantee on runtime.

5. ALGORITHM CORRECTNESS

In this section we prove Theorem 4.1, first giving a bound on the run-time in Subsection 5.1 and then showing correctness in Subsection 5.2.

5.1 Entropy increment argument and run-time bound

In this subsection we bound the size of the pseudo-neighborhood constructed, but make no guarantee that it actually contains the true neighborhood. We use several standard information-theoretic quantities including entropy, Kullback-Leibler divergence, and mutual information. The relevant definitions can be found in any information theory textbook (such as [16]). The following lemma gives a lower bound on the conditional mutual information of each added node.

Lemma 5.1. *Assume that event $\mathcal{A}(\ell,\epsilon)$ holds and suppose that* LearnNbhd *added node $j_{\ell+1}$ to the pseudo-neighborhood of u after having added $j_1,\dots,j_\ell$. Then the conditional mutual information $I(X_u;X_{j_{\ell+1}}|X_{j_1},\dots,X_{j_\ell}) \ge \frac{1}{2}(\tau-\epsilon)^2$.*

We can now argue that the number of nodes added in the pseudo-neighborhood step is not too large by using a potential argument. The bound is stated in terms of the quantity

$$\ell(\tau,\epsilon) = \frac{2}{(\tau-\epsilon)^2}\,.$$

Lemma 5.2. *If event $\mathcal{A}(\ell(\tau,\epsilon),\epsilon)$ holds, then at the end of the pseudo-neighborhood construction step the set $\mathcal{S}$ has cardinality at most $\ell(\tau,\epsilon) = 2(\tau-\epsilon)^{-2}$.*

Before proving the lemma, let us quickly see how it justifies the runtime claimed in Theorem 4.1. Each maximization step in line 2 of LearnNbhd takes time $\mathcal{O}(pn)$, so we get a cost of $\mathcal{O}(|\mathcal{S}|pn) = \mathcal{O}(\ell^* pn)$ for the pseudo-neighborhood step, and this dominates the runtime.

Proof of Lemma 5.2. Consider the sequence of nodes added, $j_1, j_2, \dots, j_r$, with $\mathcal{S} = \{j_1, j_2, \dots, j_r\}$. Then

$$\begin{aligned} 1 \ge H(X_u) &\overset{(a)}{\ge} I(X_u|X_{\mathcal{S}}) \overset{(b)}{=} \sum_{k=1}^{r} I(X_u; X_{j_k}|X_{j_1},\dots,X_{j_{k-1}}) \\ &\overset{(c)}{\ge} \min\{\ell+1, r\}\cdot\tfrac{1}{2}(\tau-\epsilon)^2\,. \end{aligned}$$

(a) is by non-negativity of conditional entropy and the definition of mutual information, $I(X_u|X_{\mathcal{S}}) = H(X_u) - H(X_u|X_{\mathcal{S}})$, (b) follows by the chain rule for mutual information, and (c) is by Lemma 5.1. Since $\ell+1$ is strictly larger than $2(\tau-\epsilon)^{-2}$, $r = |\mathcal{S}|$ must satisfy the bound stated in the lemma. □

Proof of Lemma 5.1. Let $\mathcal{S}_\ell = (j_1,\dots,j_\ell)$ consist of the first ℓ nodes already added to the pseudo-neighborhood. Our goal is to show that the next node $j_{\ell+1}$ has mutual information $I(X_u; X_{j_{\ell+1}}|X_{\mathcal{S}_\ell}) \ge \frac{1}{2}(\tau-\epsilon)^2$.

We use a shorthand for the (random) conditional measure: $Q(u+) = \mathsf{P}(X_u = +|X_{\mathcal{S}_\ell})$ and similarly $Q(u+|i-) = \mathsf{P}(X_u = +|X_i = -, X_{\mathcal{S}_\ell})$, with similar definitions for any combination of '+' and '−'. Thus we can write

$$\nu^{\mathtt{avg}}_{u|i;\mathcal{S}_\ell} = 2\cdot \mathsf{E}_{X_{\mathcal{S}_\ell}}\Big(Q(i+)Q(i-)\big|Q(u+|i+) - Q(u+|i-)\big|\Big)\,.$$

Now for any $i \in \mathcal{V}\setminus(\mathcal{S}\cup\{u\})$,

$$\begin{aligned} &\sqrt{\tfrac{1}{2}\cdot I\big(X_u; X_i|X_{\mathcal{S}_\ell}\big)} \\ &= \sqrt{\tfrac{1}{2}\cdot \sum_{x_{\mathcal{S}_\ell}} P(x_{\mathcal{S}_\ell}) I\big(X_u; X_i|X_{\mathcal{S}_\ell} = x_{\mathcal{S}_\ell}\big)} \\ &\overset{(a)}{\ge} \sum_{x_{\mathcal{S}_\ell}} P(x_{\mathcal{S}_\ell}) \sqrt{\tfrac{1}{2}\cdot I\big(X_u; X_i|X_{\mathcal{S}_\ell} = x_{\mathcal{S}_\ell}\big)} \\ &= \mathsf{E}_{X_{\mathcal{S}_\ell}} \sqrt{\tfrac{1}{2}\cdot D_{\mathrm{KL}}\big(Q(u,i)\,\|\,Q(u)Q(i)\big)} \\ &\overset{(b)}{\ge} \mathsf{E}_{X_{\mathcal{S}_\ell}} D_{\mathrm{TV}}(Q(u,i), Q(u)Q(i)) \\ &\overset{(c)}{\ge} \mathsf{E}_{X_{\mathcal{S}_\ell}} \big|Q(u+,i+) - Q(u+)Q(i+)\big| \\ &= \mathsf{E}_{X_{\mathcal{S}_\ell}} \big|Q(u+|i+)Q(i+) - Q(u+)Q(i+)\big| \\ &= \mathsf{E}_{X_{\mathcal{S}_\ell}} \Big(Q(i+)\cdot\big|Q(u+|i+)\big(Q(i+)+Q(i-)\big) \\ &\qquad - Q(u+|i+)Q(i+) - Q(u+|i-)Q(i-)\big|\Big) \\ &= \mathsf{E}_{X_{\mathcal{S}_\ell}} \Big(Q(i+)Q(i-)\cdot\big|Q(u+|i+) - Q(u+|i-)\big|\Big) \\ &= \tfrac{1}{2}\cdot \nu^{\mathtt{avg}}_{u|i;\mathcal{S}_\ell} \\ &\overset{(e)}{\ge} \tfrac{1}{2}\cdot\big(\widehat{\nu}^{\mathtt{avg}}_{u|i;\mathcal{S}_\ell} - \epsilon\big)\,. \end{aligned}$$

(a) is by Jensen's inequality applied to the (concave) square root function, (b) Pinsker's inequality, (c) the definition of total variation distance, (d) Lemma 2.1, (e) is by definition of $\mathcal{A}(\ell,\epsilon)$ and the fact that the conditioning set has cardinality $|\mathcal{S}_\ell| \le \ell$.

Finally, by definition of the algorithm, node $j_{\ell+1}$ is only added to $\mathcal{S}_\ell$ if $|\widehat{\nu}^{\mathtt{avg}}_{u|j_{\ell+1};\mathcal{S}_\ell}| \ge \tau$, so the previous displayed equation implies that $I(X_u; X_{j_{\ell+1}}|X_{\mathcal{S}_\ell}) \ge \frac{1}{2}(\tau-\epsilon)^2$ as claimed. □

5.2 Key structural result and algorithm correctness

We now state our structural result, and use it to prove correctness of the algorithm. Its proof is given in Section 6. In this subsection we use the values $\tau^*, \epsilon^*, \ell^*$ defined in Theorem 4.1.

Proposition 5.3. *Let G be a graph of maximum degree d, and consider an Ising model (1) on G with vector of parameters $\theta \in \Omega_{\alpha,\beta,h}(G)$. For any node $u \in \mathcal{V}$, if $\mathcal{S} \subseteq \mathcal{V}\setminus\{u\}$ such that $\partial u \not\subseteq \mathcal{S}$, then there exists a node $i \in \partial u \setminus \mathcal{S}$ with $\nu^{\mathtt{avg}}_{u|i;\mathcal{S}} \ge 2\tau^*$.*

We now show that the pseudo-neighborhood contains the true neighborhood.

Corollary 5.4. *If event $\mathcal{A}(\ell^*,\epsilon^*)$ holds, then for any $u \in \mathcal{V}$, the pseudo-neighborhood $\mathcal{S}$ constructed by* LearnNbhd$(X^{1:n}, \tau^*, u)$ *contains the true neighborhood ∂u.*

Proof. Consider an arbitrary node $u \in \mathcal{V}$ and suppose $\partial u \not\subseteq \mathcal{S}$. Proposition 5.3 shows that $\nu^{\mathtt{avg}}_{u|i;\mathcal{S}} \ge 2\tau^*$ for some $i \in \partial u \setminus \mathcal{S}$. If event $\mathcal{A}(\ell^*,\epsilon^*)$ holds, then $|\mathcal{S}| \le \ell^*$ by Lemma 5.2, and we have $\widehat{\nu}^{\mathtt{avg}}_{u|i;\mathcal{S}} \ge \nu^{\mathtt{avg}}_{u|i;\mathcal{S}} - \epsilon^* \ge 3\tau^*/2$. But this contradicts line 3 of algorithm LearnNbhd. □

Corollary 5.5. *Consider the same setup as Corollary 5.4. After the pruning step, $\mathcal{S} = \partial u$.*

Proof. By Corollary 5.4, the pseudo-neighborhood $\mathcal{S}$ contains ∂u, hence Equation (3) states that $\nu^{\mathtt{avg}}_{u|i;\mathcal{S}\setminus\{i\}} = 0$ for non-neighbors i. By Lemma 5.2, $|S| \le \ell^*$, and by definition of the event $\mathcal{A}(\ell^*,\epsilon^*)$ (with our choice $\epsilon^* = \tau^*/2$), we have $\widehat{\nu}^{\mathtt{avg}}_{u|i;\mathcal{S}\setminus\{i\}} \le \epsilon^* = \tau^*/2$ for all non-neighbors i, and hence these are discarded. Conversely, by Proposition 5.3, $\widehat{\nu}^{\mathtt{avg}}_{u|i;\mathcal{S}\setminus\{i\}} \ge 3\tau^*/2$, so no neighbors are discarded. □

All the ingredients are in place to finish the proof of Theorem 4.1.

Proof of Theorem 4.1. By Lemma 3.2, our choice of n in the statement of the theorem guarantees $\mathsf{P}(\mathcal{A}(\ell^*,\epsilon^*)) \ge 1-\zeta$. Together the two corollaries prove correctness of the algorithm assuming event $\mathcal{A}(\ell^*,\epsilon^*)$ holds, and this completes the proof of Theorem 4.1, modulo Proposition 5.3. □

6. PROOF OF PROPOSITION 5.3

Fix $u \in \mathcal{V}$ and $\mathcal{S} \subseteq \mathcal{V} \setminus \{u\}$, and let $\mathcal{U} = \partial u \setminus \mathcal{S}$ consist of the neighbors of u not in $\mathcal{S}$. Assume that $\mathcal{U}$ is nonempty or there is nothing to prove. Let $\theta_{u\mathcal{U}} := (\theta_{ui})_{i\in\mathcal{U}}$ and recall the definition $\tau^* := \alpha^2\delta^{4d+1}/16d\beta$. We will prove that for any assignment $x_\mathcal{S} \in \{-,+\}^\mathcal{S}$,

$$\sum_{i\in\mathcal{U}} \theta_{ui} \cdot \lambda_i(x_\mathcal{S})\nu_{u|i;x_\mathcal{S}} \geq \|\theta_{u\mathcal{U}}\|_1 \cdot 2\tau^* \,. \tag{4}$$

Averaging with respect to $x_\mathcal{S}$ and applying the triangle inequality gives

$$\sum_{i\in\mathcal{U}} \theta_{ui} \cdot \nu^{\mathtt{avg}}_{u|i;\mathcal{S}} = \sum_{i\in\mathcal{U}} \theta_{ui} \cdot \mathsf{E}\big(\lambda_i(X_\mathcal{S})|\nu_{u|i;X_\mathcal{S}}|\big) \geq \|\theta_{u\mathcal{U}}\|_1 \cdot 2\tau^* \,.$$

As a consequence there exists some $i \in \mathcal{U}$ such that

$$\nu^{\mathtt{avg}}_{u|i;\mathcal{S}} \geq 2\tau^* \,,$$

and this proves the Proposition.

We proceed with showing (4). Let $\widetilde{\mathcal{S}} = \partial u \cap \mathcal{S}$ consist of nodes in $\mathcal{S}$ adjacent to u and $\widetilde{\theta_u} = \theta_u + \sum_{j\in\widetilde{\mathcal{S}}} \theta_{uj}x_j$ be the effective external field at u when we include the effect due to $x_{\widetilde{\mathcal{S}}}$. Using the notation $Q(\cdot)$ for the conditional measure $\mathsf{P}(\,\cdot\,|X_\mathcal{S} = x_\mathcal{S})$, and $Q(u+|x_\mathcal{U}) = \mathsf{P}(X_u = +|X_\mathcal{U} = x_\mathcal{U}, X_\mathcal{S} = x_\mathcal{S})$, we let

$$\begin{aligned} g(x_\mathcal{U}) &:= Q(u+|x_\mathcal{U}) \\ &= \frac{\exp\{2(\theta_u + \sum_{j\in\mathcal{U}} \theta_{uj}x_j + \sum_{j\in\widetilde{\mathcal{S}}} \theta_{uj}x_j)\}}{1+\exp\{2(\theta_u + \sum_{j\in\mathcal{U}} \theta_{uj}x_j + \sum_{j\in\widetilde{\mathcal{S}}} \theta_{uj}x_j)\}} \\ &= \frac{\exp\{2(\widetilde{\theta_u} + \sum_{j\in\mathcal{U}} \theta_{uj}x_j)\}}{1+\exp\{2(\widetilde{\theta_u} + \sum_{j\in\mathcal{U}} \theta_{uj}x_j)\}} \,. \end{aligned} \tag{5}$$

Suppose $i \in \mathcal{U}$. Conditioning on the values of the remaining neighbors $\mathcal{U}\setminus\{i\}$ of u,

$$\begin{aligned} \nu_{u|i;x_\mathcal{S}} &= Q(u+|i+) - Q(u+|i-) \\ &= \sum_{x_{\mathcal{U}\setminus\{i\}}} \Big(Q(u+|i+, x_{\mathcal{U}\setminus\{i\}})Q(x_{\mathcal{U}\setminus\{i\}}|i+) \\ &\qquad - Q(u+|i-, x_{\mathcal{U}\setminus\{i\}})Q(x_{\mathcal{U}\setminus\{i\}}|i-)\Big) \qquad (6) \\ &= \sum_{x_\mathcal{U}} \left(Q(u+|x_\mathcal{U})Q(x_\mathcal{U}) \frac{\mathbb{1}_{\{x_i=+\}}}{Q(i+)} \right. \qquad (7) \\ &\qquad \left. - Q(u+|x_\mathcal{U})Q(x_\mathcal{U}) \frac{\mathbb{1}_{\{x_i=-\}}}{Q(i-)} \right) \\ &= \mathsf{E}\Big(g(Y) \frac{Y_i}{Q(Y_i)} \Big) \,. \qquad (8) \end{aligned}$$

The expectation is over the random vector $Y \in \{-,+\}^{|\mathcal{U}|}$ with law $\mathsf{P}(Y \in \cdot) = \mathsf{P}(X_\mathcal{U} \in \cdot | X_\mathcal{S} = x_\mathcal{S})$, and in particular $\mathsf{P}(Y_i = +) = \mathsf{P}(X_i = +|X_\mathcal{S} = x_\mathcal{S}) = Q(i+)$. The notation $Q(Y_i)$ is understood to mean $Q(i+)$ if $Y_i = +$ and $Q(i-)$ if $Y_i = -$.

It is helpful to rescale and shift the variable $Y_i/Q(Y_i)$ in (8) so that it takes values ± 1. To this end, define the quantities

$$s_i := \frac{1}{2} \cdot \left(\frac{1}{Q(i+)} - \frac{1}{1-Q(i+)} \right) \quad \text{and}$$
$$t_i := \frac{1}{\lambda_i(x_\mathcal{S})} = \frac{1}{2} \cdot \left(\frac{1}{Q(i+)} + \frac{1}{1-Q(i+)} \right) .$$

(The function $\lambda_i(x_\mathcal{S})$ was defined at the beginning of Section 3.) Arithmetic manipulations lead to

$$\mathsf{E}Y_i = 2Q(i+) - 1 = -\frac{s_i}{t_i} \qquad \text{and}$$
$$Y_i = \frac{Y_i}{t_iQ(Y_i)} - \frac{s_i}{t_i} = \frac{Y_i}{t_iQ(Y_i)} + \mathsf{E}Y_i \,.$$

Multiplying (8) by $2\theta_{ui}/t_i$ and using the identities in the last display gives

$$\begin{aligned} 2\frac{\theta_{ui}}{t_i} \cdot \nu_{u|i;x_\mathcal{S}} &= \mathsf{E}\Big(2g(Y)\frac{Y_i\theta_{ui}}{t_iq(y_i)}\Big) \\ &= \mathsf{E}\Big(2g(Y)\big(Y_i\theta_{ui} - \mathsf{E}Y_i\theta_{ui}\big)\Big) \\ &= \mathsf{E}\Big((2g(Y)-1)(Y_i\theta_{ui} - \mathsf{E}Y_i\theta_{ui})\Big) \,. \end{aligned}$$

Summing the last displayed quantity over $i \in \mathcal{U}$ gives

$$\begin{aligned} &\sum_{i\in\mathcal{U}} \theta_{ui} \cdot \lambda_i(x_\mathcal{S})\nu_{u|i;x_\mathcal{S}} \\ &= \tfrac{1}{2} \cdot \mathsf{E}\Big((2g(Y)-1)(Y\cdot\theta_{u\mathcal{U}} - \mathsf{E}Y\cdot\theta_{u\mathcal{U}})\Big) \\ &\geq d\beta 2\tau^* \geq \|\theta_{u\mathcal{U}}\|_1 \cdot 2\tau^* \,. \end{aligned} \tag{9}$$

Here $\theta_{u\mathcal{U}} = (\theta_{ui})_{i\in\mathcal{U}}$ and the first inequality is by Lemma 7.3, given in Section 7 below. □

7. TECHNICAL LEMMA

The goal of this section is to justify Equation (9), which appeared in the proof of Proposition 5.3. We start with an observation that will be used in the proof. Recall that Y is a random vector equal in distribution to $X_\mathcal{U}$ conditioned on $X_\mathcal{S} = x_\mathcal{S}$. Due to the "conditional randomness" Lemma 2.1, Y has probability at least $\delta^{|\mathcal{U}|}$ of taking each value in $\{-,+\}^{|\mathcal{U}|}$, where $\delta = \frac{1}{2}\exp^{-2(\beta d+h)}$. Hence we can decompose the probability mass function P_Y of Y as

$$P_Y(y) = \delta^{|\mathcal{U}|} + \overline{P}_Y(y) \tag{10}$$

with $\overline{P}_Y(y) \geq 0$ for all $y \in \{-,+\}^{|\mathcal{U}|}$. We will be concerned with the random variable

$$Z := Y \cdot \theta_{u\mathcal{U}} + \widetilde{\theta_u} \,, \tag{11}$$

and the decomposition (10) will allow us to obtain anti-concentration for Z from anti-concentration for sums of i.i.d. uniform ± 1 random variables.

The following result of Erdös on the Littlewood-Offord problem shows anti-concentration for weighted sums of i.i.d. uniform ± 1 random variables. (It can be found, e.g., as Corollary 7.4 in [68] and is a simple consequence of Sperner's Lemma).

Lemma 7.1 (Erdös [25])**.** *Let $w_1, \ldots, w_r$ be real numbers with $|w_i| \geq \alpha$ for all i. Let $I = \{t \in \mathbb{R} : t_0 - \alpha < t < t_0 + \alpha\}$ be an open interval of length 2α. If $\xi = (\xi_1, \ldots, \xi_r)$ is uniformly distributed on $\{-1,1\}^r$, then*

$$\mathsf{P}(w\cdot\xi \in I) \leq \frac{1}{2^r} \cdot \binom{r}{\lfloor \frac{r}{2} \rfloor} \leq \frac{1}{2} \,.$$

We can use the decomposition (10) and Lemma 7.1 to draw the following conclusion. (It is possible to show this directly, but this approach seems clearer.) We mention that the only place the lower bound α on the coupling strengths appears is through the following lemma.

Lemma 7.2. *Let $2t = 2^{|\mathcal{U}|} \cdot \delta^{|\mathcal{U}|}$. Consider the random variable Z defined in* (11)*, and let $\mu = \mathsf{E}Z$. Then*

$$\mathsf{E}[(\mu - Z)\mathbb{1}_{\{Z \le \mu - \alpha t\}}] \ge \frac{\alpha t^2}{2}\,.$$

Proof. Decompose the probability mass function P_Y as discussed at the beginning of this section in (10). Let $2t = 2^{|\mathcal{U}|}\cdot\delta^{|\mathcal{U}|}$ be the total mass assigned to the uniform part. Let $Z' = \xi\cdot\theta_{u\mathcal{U}} + \widetilde{\theta_u}$ where $\xi \in \{-,+\}^{|\mathcal{U}|}$ is uniformly distributed and let $Z'' = Y''\cdot\theta_{u\mathcal{U}} + \widetilde{\theta_u}$ where $Y'' \sim (1-M)^{-1}\overline{P}_Y$. The variable Z can be represented as a mixture distribution: if we define $R \sim \mathrm{Ber}(2t)$, then

$$Z \stackrel{d}{=} R\cdot Z' + (1-R)\cdot Z''\,,$$

and we can think of obtaining Z by choosing either Z' or Z'' with probabilities $2t$ or $(1-2t)$. Let $I_t = (\mu - t\alpha, \mu + (2-t)\alpha)$. Lemma 7.1 implies that $\mathsf{P}(Z' \notin I_t) \ge 1/2$ for any t, and hence $\mathsf{P}(Z \notin I_t) \ge t$. Denote the probability that Z lies to the left of I_t and inside I_t, respectively, as

$$m_1 = \mathsf{P}(Z \le \mu - t\alpha) \qquad \text{and} \qquad m_2 = \mathsf{P}(Z \in I_t) \le 1 - t\,,$$

so that $1 - m_1 - m_2 = \mathsf{P}(Z \ge \mu + (2-t)\alpha)$. Thinking about placing the probability mass to minimize μ subject to fixed m_1 and m_2 justifies the inequality

$$\begin{aligned}\mu = \mathsf{E}Z &= m_1\cdot\mathsf{E}(Z \mid Z \le \mu - t\alpha) + m_2\cdot\mathsf{E}(Z \mid Z \in I_t)\\ &\quad + (1-m_1-m_2)\mathsf{E}(Z \mid Z \ge \mu + (2-t)\alpha)\\ &\ge m_1\cdot\mathsf{E}(Z \mid Z \le \mu - t\alpha) + m_2(\mu - t\alpha)\\ &\quad + (1-m_1-m_2)(\mu + (2-t)\alpha)\,.\end{aligned}$$

Using $m_2 \le 1 - t$ and performing arithmetic manipulations leads to

$$\begin{aligned}m_1\cdot\mathsf{E}[(\mu - \alpha t) - Z \mid Z \le \mu - \alpha t] + m_1 2\alpha&\\ \ge -m_2 2\alpha + (2-t)\alpha \ge t\alpha\,.&\end{aligned}$$

At least one of the two terms on the left-hand side is larger than the average, which is at least $t\alpha/2$, and in either case $\mathsf{E}[(\mu - Z)\mathbb{1}_{\{Z\le\mu-\alpha t\}}] \ge \alpha t^2/2$ (using the fact that $t \le 1$). □

The remainder of this section is devoted to proving the following lemma.

Lemma 7.3. *Let g and Y be as in the proof of Proposition 5.3 in Section 6. Then the quantity in Equation* (9) *is lower bounded as*

$$\tfrac{1}{2}\mathsf{E}\Big((2g(Y)-1)(Y\cdot\theta_{u\mathcal{U}} - \mathsf{E}Y\cdot\theta_{u\mathcal{U}})\Big) \ge \frac{\alpha^2\delta^{4|\mathcal{U}|+1}}{8} \ge d\beta 2\tau^*\,.$$

Proof. We start by adding and subtracting $\widetilde{\theta_u}$ to the left-hand side of the lemma statement:

$$\begin{aligned}&\mathsf{E}\Big(\big(2g(Y)-1\big)\big(Y\cdot\theta_{u\mathcal{U}} - \mathsf{E}Y\cdot\theta_{u\mathcal{U}}\big)\Big)\\ &= \mathsf{E}\Big(\big(2g(Y)-1\big)\big(Y\cdot\theta_{u\mathcal{U}} + \widetilde{\theta_u} - \mathsf{E}(Y\cdot\theta_{u\mathcal{U}} + \widetilde{\theta_u})\big)\Big)\,.\end{aligned}$$

Recalling the definition of g in (5), we make the observation that $2g(x) - 1 = \tanh(x\cdot\theta_{u\mathcal{U}} + \widetilde{\theta_u})$. We will use the fact that $\tanh(z)$ is an odd, increasing function, which is concave for $t \ge 0$ and convex for $t \le 0$. Recall from (11) the definition $Z = Y\cdot\theta_{u\mathcal{U}} + \widetilde{\theta_u}$ and let $\mu := \mathsf{E}Z$. The lemma statement requires that we lower bound $\mathsf{E}[\tanh(Z)(Z-\mu)]$. We assume from now onward that $\mu \ge 0$, but a (symmetrically) identical argument applies to the opposite case $\mu \le 0$. From the definition of Z and assumptions $|\theta_{ui}| \le \beta$, $|\theta_u| \le h$ we obtain the bound

$$\mu = \mathsf{E}Z \le (|\mathcal{U}| + |\widetilde{\mathcal{S}}|)\cdot\beta + h \le \beta d + h\,.$$

Next we record a few estimates on the function $\tanh(\cdot)$. The derivative satisfies

$$\frac{d}{dz}\tanh(z) = \frac{4}{(e^z + e^{-z})^2} \ge \frac{1}{e^{2|z|}}$$

and due to the concavity of $\tanh(z)$ for $z \ge 0$, we have the estimate

$$\begin{aligned}\tanh(z) &\le \tanh(\mu) - \frac{(\mu - z)}{e^{2\mu}} \le \tanh(\mu) - \frac{(\mu - z)}{e^{2(\beta d + h)}}\\ &= \tanh(\mu) - 2\delta(\mu - z) \qquad \text{for } 0 \le z \le \mu\,.\end{aligned}$$

We additionally use the bound $\tanh(t) \ge \tanh(\mu)$ for $z \ge \mu$ due to monotonicity of $\tanh(\cdot)$. Partitioning the range of Z and using these estimates gives

$$\begin{aligned}&\mathsf{E}[\tanh(Z)(Z-\mu)]\\ &= \mathsf{E}[\tanh(Z)(Z-\mu)\mathbb{1}_{\{Z<0\}}] + \mathsf{E}[\tanh(Z)(Z-\mu)\mathbb{1}_{\{Z\in[0,\mu]\}}]\\ &\quad + \mathsf{E}[\tanh(Z)(Z-\mu)\mathbb{1}_{\{Z>\mu\}}]\\ &\ge \mathsf{E}[\tanh(Z)(Z-\mu)\mathbb{1}_{\{Z<0\}}]\\ &\quad + \mathsf{E}\Big[\Big(\tanh(\mu) - 2\delta(\mu - Z)\Big)(Z-\mu)\mathbb{1}_{\{Z\in[0,\mu]\}}\Big]\\ &\quad + \mathsf{E}[\tanh(\mu)(Z-\mu)\mathbb{1}_{\{Z>\mu\}}]\,.\end{aligned}$$

Subtracting $\tanh(\mu)\mathsf{E}(Z-\mu) = 0$ from the third term, the last expression is equal to

$$\begin{aligned}&\mathsf{E}[\tanh(Z)(Z-\mu)\mathbb{1}_{\{Z<0\}}]\\ &\quad + \mathsf{E}\Big[\Big(\tanh(\mu) - 2\delta(\mu - Z)\Big)(Z-\mu)\mathbb{1}_{\{Z\in[0,\mu]\}}\Big]\\ &\quad - \mathsf{E}[\tanh(\mu)(Z-\mu)\mathbb{1}_{\{Z<0\}}]\\ &\quad - \mathsf{E}[\tanh(\mu)(Z-\mu)\mathbb{1}_{\{Z\in[0,\mu]\}}]\\ =&\mathsf{E}[\big(\tanh(Z) - \tanh(\mu)\big)(Z-\mu)\mathbb{1}_{\{Z<0\}}]\\ &\quad + \mathsf{E}\Big[2\delta(\mu - Z)^2\mathbb{1}_{\{Z\in[0,\mu]\}}\Big]\,.\end{aligned} \tag{12}$$

Both of these terms are non-negative.

Lemma 7.2 states that $\mathsf{E}[(\mu - Z)\mathbb{1}_{\{Z\le\mu-\alpha t\}}] \ge \frac{\alpha t^2}{2}$, where $2t = (2\delta)^{|\mathcal{U}|}$, and this means that either

$$\begin{aligned}\mathsf{E}[(\mu - Z)\mathbb{1}_{\{Z\le\mu-\alpha t\}}\mathbb{1}_{\{Z<0\}}] \ge \frac{\alpha t^2}{4} \qquad &\text{or}\\ \mathsf{E}[(\mu - Z)\mathbb{1}_{\{Z\le\mu-\alpha t\}}\mathbb{1}_{\{Z\in[0,\mu]\}}] \ge \frac{\alpha t^2}{4}\,,&\end{aligned} \tag{13}$$

(or both) is true. In the former case, the first term in (12) is lower bounded by

$$\begin{aligned}&\mathsf{E}[\big(\tanh(Z) - \tanh(\mu)\big)(Z-\mu)\mathbb{1}_{\{Z\le\mu-\alpha t\}}\mathbb{1}_{\{Z<0\}}]\\ &\ge \mathbb{1}_{\{\mu<1\}}\frac{\alpha t^2}{4}\big(\tanh(\mu) - \tanh(\mu - \alpha t)\big) + \mathbb{1}_{\{\mu\ge1\}}\frac{\alpha t^2}{4}\tanh(1)\end{aligned} \tag{14}$$

$$\stackrel{(a)}{\ge} \frac{\alpha t^2}{4}\min\Big(\frac{\alpha t}{e^2}, \tanh(1)\Big) \stackrel{(b)}{\ge} \frac{\alpha^2 t^3}{4e^2} \ge \frac{\alpha^2 t^3}{2^5}\,. \tag{15}$$

(a) follows by bounding the first term in (14) for $0 \le \mu < 1$ by noting that $[\mu - \alpha t, \mu] \subseteq [-\alpha t, 1] \subseteq [-1, 1]$ and lower bounding the derivative of $\tanh(z)$ on $[-1,1]$ by $1/e^2$, and (b) follows from the fact that $\tanh(1) > 1/e^2 \ge \alpha t/e^2$.

In the latter case of (13), the second term of (12) is lower bounded by

$$2\delta \cdot \mathsf{E}\Big((\mu - Z)^2 \mathbb{1}_{\{Z \le \mu - \alpha t\}} \mathbb{1}_{\{Z \in [0,\mu]\}}\Big) \ge \frac{2\delta\alpha^2 t^4}{16}. \quad (16)$$

This used Cauchy-Schwarz (or equivalently non-negativity of the variance) to lower bound the expectation of $(\mu - Z)^2$ by $(\alpha t^2/4)^2$, the square of the expectation given in (13). The quantity in (16) is smaller than in (15), because $2\delta \le 1$ and $t \le 1/2$. Multiplying the right-hand side of (16) by $1/2$ and plugging in $t = \frac{1}{2}(2\delta)^{|\mathcal{U}|} \ge \delta^{|\mathcal{U}|}$ completes the proof of the lemma. □

8. DISCUSSION

Our algorithm learns Ising models in time quadratic in the number of nodes (ignoring the log factor). In light of Valiant's [71] algorithm for finding large correlations in less than quadratic time, it is plausible that one could achieve an analogous further improvement of the runtime to p^c for some constant $c < 2$. Perhaps even "input-sparsity" time $\mathcal{O}(pd)$ is possible. As far as practical applicability, it seems most urgent to improve upon the doubly-exponential dependence of the run-time and sample complexity on βd. Many generalizations and extensions of the results are likely possible, including to pairwise Markov random fields with alphabet sizes larger than two.

9. ACKNOWLEDGEMENTS

I am extremely grateful to David Gamarnik and Devavrat Shah for countless discussions on graphical models and related topics over the last two years. I thank Bruce Hajek for comments on a draft of the paper and Sahand Negahban and Costis Daskalakis for stimulating conversations. This work was supported in part by NSF grants CMMI-1462158, CMMI-1335155, and CNS-1161964, and by Army Research Office MURI Award W911NF-11-1-0036.

10. REFERENCES

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

[2] D. Ackley, G. Hinton, and T. Sejnowski. A learning algorithm for boltzmann machines*. *Cognitive science*, 9(1):147–169, 1985.

[3] A. Anandkumar, F. Huang, D. Hsu, and S. Kakade. Learning mixtures of tree graphical models. In *NIPS*, 2012.

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

[5] E. Aurell, C. Ollion, and Y. Roudi. Dynamics and performance of susceptibility propagation on synthetic data. *The European Physical Journal B-Condensed Matter and Complex Systems*, 77(4):587–595, 2010.

[6] A. Bandyopadhyay and D. Gamarnik. Counting without sampling: Asymptotics of the log-partition function for certain statistical physics models. *Random Structures & Algorithms*, 33(4):452–479, 2008.

[7] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

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

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

[10] G. Bresler, D. Gamarnik, and D. Shah. Structure learning of antiferromagnetic Ising models. In *NIPS*, 2014.

[11] G. Bresler, E. Mossel, and A. Sly. Reconstruction of Markov random fields from samples: Some observations and algorithms. In *APPROX*, 2008.

[12] S. Brush. History of the Lenz-Ising Model. *Rev. Mod. Phys.*, 39(4):883–893, Oct 1967.

[13] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory*, 14(3):462–467, 1968.

[14] S. Cocco, S. Leibler, and R. Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences*, 106(33):14058–14062, 2009.

[15] S. Cocco and R. Monasson. Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. *Journal of Statistical Physics*, pages 1–63, 2012.

[16] T. Cover and J. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[17] I. Csiszár and Z. Talata. Consistent estimation of the basic neighborhood of Markov random fields. *Annals of Stat.*, pages 123–145, 2006.

[18] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *Information Theory, IEEE Transactions on*, 55(5):2230–2249, 2009.

[19] S. Dasgupta. Learning polytrees. In *UAI*, 1999.

[20] A. Decelle and F. Ricci-Tersenghi. Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of ising models. *Physical review letters*, 112(7):070603, 2014.

[21] I. Diakonikolas and R. A. Servedio. Improved approximation of linear threshold functions. *computational complexity*, 22(3):623–677, 2013.

[22] R. Dobrushin. Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3):458–486, 1970.

[23] R. Dobrushin and S. Shlosman. Constructive criterion for the uniqueness of Gibbs field. In *Statistical physics and dynamical systems*, pages 347–370. Springer, 1985.

[24] M. Dyer, A. Sinclair, E. Vigoda, and D. Weitz. Mixing in time and space for lattice spin systems: A combinatorial view. *Random Structures & Algorithms*, 24(4):461–479, 2004.

[25] P. Erdös. On a lemma of Littlewood and Offord. *Bulletin of the American Mathematical Society*, 51(12):898–902, 1945.

[26] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[27] J. Friedman, T. Hastie, R. Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

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

[29] D. Gamarnik and D. Katz. Correlation decay and deterministic fptas for counting list-colorings of a graph. In *SODA*, pages 1245–1254, 2007.

[30] L. Goldberg, M. Jerrum, and M. Paterson. The computational complexity of two-state spin systems. *Random Structures & Algorithms*, 23(2):133–154, 2003.

[31] N. Goyal, S. Vempala, and Y. Xiao. Fourier pca. In *Symposium on the Theory of Computing (STOC)*, page 3, 2014.

[32] G. Hinton and T. Sejnowski. Learning and relearning in Boltzmann machines. *MIT Press*, 1(282-317):4–2, 1986.

[33] E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[34] A. Jalali, C. Johnson, and P. Ravikumar. On learning discrete graphical models using greedy methods. *arXiv preprint arXiv:1107.3258*, 2011.

[35] A. Jalali, P. Ravikumar, V. Vasuki, and S. Sanghavi. On learning discrete graphical models using group-sparse regularization. In *AISTATS*, volume 14, 2011.

[36] M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993.

[37] A. Kalai, A. Samorodnitsky, and S. Teng. Learning and smoothed analysis. In *FOCS*, pages 395–404. IEEE, 2009.

[38] D. Kane, A. Klivans, and R. Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In *Conference on Learning Theory*, pages 522–545, 2013.

[39] A. Klivans and R. Meka. Moment-matching polynomials. *arXiv preprint arXiv:1301.0820*, 2013.

[40] S. Lauritzen. *Graphical models*. Oxford University Press, 1996.

[41] S. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of Markov networks using ℓ_1-regularization. In *NIPS*, pages 817–824, 2006.

[42] T. Lezon, J. Banavar, M. Cieplak, A. Maritan, and N. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences*, 103(50):19033–19038, 2006.

[43] F. Martinelli and E. Olivieri. Approach to equilibrium of Glauber dynamics in the one phase region. *Comm. in Mathematical Physics*, 161(3):447–486, 1994.

[44] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[45] M. Mézard and T. Mora. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris*, 103(1):107–113, 2009.

[46] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

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

[48] T. Mora, A. Walczak, W. Bialek, and C. Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410, 2010.

[49] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[50] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[51] P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai. Greedy learning of Markov network structure. In *48th Allerton Conference*, pages 1295–1302, 2010.

[52] R. Niedermeier. *Invitation to fixed-parameter algorithms*. Oxford University Press, 2006.

[53] R. O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[54] P. Ravikumar, M. Wainwright, and J. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

[55] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[56] A. Ray, S. Sanghavi, and S. Shakkottai. Greedy learning of graphical models with small girth. In *50th Allerton Conference*, 2012.

[57] F. Ricci-Tersenghi. The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. *Journal of Statistical Mechanics: Theory and Experiment*, (08), 2012.

[58] Y. Roudi, E. Aurell, and J. Hertz. Statistical physics of pairwise probability models. *Frontiers in computational neuroscience*, 3, 2009.

[59] S. Salas and A. Sokal. Absence of phase transition for antiferromagnetic Potts models via the Dobrushin uniqueness theorem. *Journal of Statistical Physics*, 86(3-4):551–579, 1997.

[60] N. P. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. on Info. Theory*, 58(7):4117–4134, 2012.

[61] E. Schneidman, M. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[62] V. Sessak and R. Monasson. Small-correlation expansions for the inverse Ising problem. *Journal of Physics A: Mathematical and Theoretical*, 42(5):055001, 2009.

[63] A. Sinclair, P. Srivastava, and M. Thurley. Approximation algorithms for two-state anti-ferromagnetic spin systems on bounded degree graphs. *Journal of Statistical Physics*, 155(4):666–686, 2014.

[64] A. Sly. Computational transition at the uniqueness threshold. In *FOCS*, pages 287–296, 2010.

[65] A. Sly and N. Sun. The computational hardness of counting in two-spin models on d-regular graphs. In *FOCS*, pages 361–369. IEEE, 2012.

[66] N. Srebro. Maximum likelihood bounded tree-width Markov networks. In *UAI*, 2001.

[67] D. Stroock and B. Zegarlinski. The logarithmic Sobolev inequality for discrete spin systems on a lattice. *Comm. in Mathematical Physics*, 149(1):175–193, 1992.

[68] T. Tao and V. Vu. *Additive combinatorics*, volume 105. Cambridge University Press, 2006.

[69] V. Temlyakov. *Greedy approximation*. Cambridge University Press, 2011.

[70] T. Toshiyuki. Mean-field theory of boltzmann machine learning. *Physical Review E*, 58(2):2302, 1998.

[71] G. Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *FOCS*, pages 11–20, 2012.

[72] L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

[73] M. Weigt, R. White, H. Szurmant, J. Hoch, and T. Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[74] D. Weitz. Counting independent sets up to the tree threshold. In *STOC*, 2006.

[75] R. Wu, R. Srikant, and J. Ni. Learning loosely connected markov random fields. *Stochastic Systems*, 3(2):362–404, 2013.

[76] T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *Information Theory, IEEE Transactions on*, 57(7):4689–4708, 2011.

APPENDIX

A. PROOF OF LEMMA 3.2

Azuma's inequality states that if $Y \sim \mathrm{Bin}(n,\mu)$, then

$$P(|Y-n\mu| > \gamma n) \le 2\exp(-2\gamma^2 n)\,,$$

so for any subset of nodes $\mathcal{W} \subseteq \mathcal{V}$ and configuration $x_{\mathcal{W}} \in \{-,+\}^{|\mathcal{W}|}$ we have

$$\mathsf{P}\Big(\big|\widehat{\mathsf{P}}(X_{\mathcal{W}}=x_{\mathcal{W}}) - \mathsf{P}(X_{\mathcal{W}}=x_{\mathcal{W}})\big| \ge \gamma\Big) \le 2\exp(-2\gamma^2 n). \tag{17}$$

There are $2^{|\mathcal{W}|}\binom{p}{|\mathcal{W}|} \le (2p)^{|\mathcal{W}|}$ such choices of $\mathcal{W}$ and $x_{\mathcal{W}}$ of a given cardinality, and hence at most $(\ell+2)(2p)^{\ell+2}$ choices of $\mathcal{W}$ and $x_{\mathcal{W}}$ with $|\mathcal{W}| \le \ell+2$.

Suppose $n \ge (2\gamma^2)^{-1}\log\big(2(\ell+2)(2p)^{\ell+2}/\zeta\big)$. An application of the union bound implies that with probability at least

$$1-(\ell+2)(2p)^{\ell+2}\cdot 2\exp(-2\gamma^2 n) \ge 1-\zeta$$

it holds that

$$\big|\widehat{\mathsf{P}}(X_{\mathcal{W}}=x_{\mathcal{W}}) - \mathsf{P}(X_{\mathcal{W}}=x_{\mathcal{W}})\big| \le \gamma \tag{18}$$

for all $\mathcal{W}$ and $x_{\mathcal{W}}$ with $|\mathcal{W}| \le \ell+2$. For the remainder of the proof assume (18) holds.

Our goal is to bound the quantity

$$\begin{aligned}&\big|\nu^{\mathtt{avg}}_{u|i;\mathcal{S}} - \widehat{\nu}^{\mathtt{avg}}_{u|i;\mathcal{S}}\big| \\ &\quad= \big|\mathsf{E}_{X_{\mathcal{S}}\sim\mathsf{P}}\big(\lambda_i(X_{\mathcal{S}})|\nu_{u|i;X_{\mathcal{S}}}|\big) - \mathsf{E}_{X_{\mathcal{S}}\sim\widehat{\mathsf{P}}}\big(\widehat{\lambda}_i(X_{\mathcal{S}})|\widehat{\nu}_{u|i;X_{\mathcal{S}}}|\big)\big|\,.\end{aligned}$$

The triangle inequality and the inequality $\big||s|-|t|\big| \le |s-t|$ for real-valued s and t gives

$$\begin{aligned}&\big|\mathsf{E}_{X_{\mathcal{S}}\sim\mathsf{P}}\big(\lambda_i(X_{\mathcal{S}})|\nu_{u|i;X_{\mathcal{S}}}|\big) - \mathsf{E}_{X_{\mathcal{S}}\sim\widehat{\mathsf{P}}}\big(\widehat{\lambda}_i(X_{\mathcal{S}})|\widehat{\nu}_{u|i;X_{\mathcal{S}}}|\big)\big| \\ &= \Big|\sum_{x_{\mathcal{S}}}\Big[\mathsf{P}(X_{\mathcal{S}}=x_{\mathcal{S}})\lambda_i(x_{\mathcal{S}})|\nu_{u|i;x_{\mathcal{S}}}| \\ &\qquad\qquad - \widehat{\mathsf{P}}(X_{\mathcal{S}}=x_{\mathcal{S}})\widehat{\lambda}_i(x_{\mathcal{S}})|\widehat{\nu}_{u|i;x_{\mathcal{S}}}|\Big]\Big| \\ &\le \sum_{x_{\mathcal{S}}}\Big|\mathsf{P}(X_{\mathcal{S}}=x_{\mathcal{S}})\lambda_i(x_{\mathcal{S}})|\nu_{u|i;x_{\mathcal{S}}}| \\ &\qquad\qquad - \widehat{\mathsf{P}}(X_{\mathcal{S}}=x_{\mathcal{S}})\widehat{\lambda}_i(x_{\mathcal{S}})|\widehat{\nu}_{u|i;x_{\mathcal{S}}}|\Big| \\ &\le \sum_{x_{\mathcal{S}}}\Big|\mathsf{P}(X_{\mathcal{S}}=x_{\mathcal{S}})\lambda_i(x_{\mathcal{S}})\nu_{u|i;x_{\mathcal{S}}} \\ &\qquad\qquad - \widehat{\mathsf{P}}(X_{\mathcal{S}}=x_{\mathcal{S}})\widehat{\lambda}_i(x_{\mathcal{S}})\widehat{\nu}_{u|i;x_{\mathcal{S}}}\Big|\,.\end{aligned}$$

Writing out the definition of $\nu_{u|i;X_{\mathcal{S}}}$ and $\widehat{\nu}_{u|i;X_{\mathcal{S}}}$, the above sum is equal to

$$\begin{aligned}&\sum_{X_{\mathcal{S}}}\Big|\mathsf{P}(X_{\mathcal{S}}=x_{\mathcal{S}})\lambda_i(x_{\mathcal{S}})\big(\mathsf{P}(X_u=+|X_i=+,X_{\mathcal{S}}=x_{\mathcal{S}}) \\ &\qquad - \mathsf{P}(X_u=+|X_i=-,X_{\mathcal{S}}=x_{\mathcal{S}})\big) \\ &\qquad - \widehat{\mathsf{P}}(X_{\mathcal{S}}=x_{\mathcal{S}})\widehat{\lambda}_i(x_{\mathcal{S}})\big(\widehat{\mathsf{P}}(X_u=+|X_i=+,X_{\mathcal{S}}=x_{\mathcal{S}}) \\ &\qquad - \widehat{\mathsf{P}}(X_u=+|X_i=-,X_{\mathcal{S}}=x_{\mathcal{S}})\big)\Big| \\ &\overset{(a)}{=} \sum_{x_{\mathcal{S}}}\Big|\Big[\lambda_i(x_{\mathcal{S}})\frac{\mathsf{P}(X_u=+,X_i=+,X_{\mathcal{S}}=x_{\mathcal{S}})}{\mathsf{P}(X_i=+|X_{\mathcal{S}}=x_{\mathcal{S}})} \\ &\qquad\qquad - \widehat{\lambda}_i(x_{\mathcal{S}})\frac{\widehat{\mathsf{P}}(X_u=+,X_i=+,X_{\mathcal{S}}=x_{\mathcal{S}})}{\widehat{\mathsf{P}}(X_i=+|X_{\mathcal{S}}=x_{\mathcal{S}})}\Big] \\ &\qquad - \Big[\lambda_i(x_{\mathcal{S}})\frac{\mathsf{P}(X_u=+,X_i=-,X_{\mathcal{S}}=x_{\mathcal{S}})}{\mathsf{P}(X_i=-|X_{\mathcal{S}}=x_{\mathcal{S}})} \\ &\qquad\qquad - \widehat{\lambda}_i(x_{\mathcal{S}})\frac{\widehat{\mathsf{P}}(X_u=+,X_i=-,X_{\mathcal{S}}=x_{\mathcal{S}})}{\widehat{\mathsf{P}}(X_i=-|X_{\mathcal{S}}=x_{\mathcal{S}})}\Big]\Big| \\ &\overset{(b)}{\le} \sum_{x_{\mathcal{S}}}\Big|\lambda_i(x_{\mathcal{S}})\frac{\mathsf{P}(X_u=+,X_i=+,X_{\mathcal{S}}=x_{\mathcal{S}})}{\mathsf{P}(X_i=+|X_{\mathcal{S}}=x_{\mathcal{S}})} \\ &\qquad\qquad - \widehat{\lambda}_i(x_{\mathcal{S}})\frac{\widehat{\mathsf{P}}(X_u=+,X_i=+,X_{\mathcal{S}}=x_{\mathcal{S}})}{\widehat{\mathsf{P}}(X_i=+|X_{\mathcal{S}}=x_{\mathcal{S}})}\Big| \\ &\quad + \sum_{x_{\mathcal{S}}}\Big|\lambda_i(x_{\mathcal{S}})\frac{\mathsf{P}(X_u=+,X_i=-,X_{\mathcal{S}}=x_{\mathcal{S}})}{\mathsf{P}(X_i=-|X_{\mathcal{S}}=x_{\mathcal{S}})} \\ &\qquad\qquad - \widehat{\lambda}_i(x_{\mathcal{S}})\frac{\widehat{\mathsf{P}}(X_u=+,X_i=-,X_{\mathcal{S}}=x_{\mathcal{S}})}{\widehat{\mathsf{P}}(X_i=-|X_{\mathcal{S}}=x_{\mathcal{S}})}\Big| \\ &:= C^+ + C^-\,.\end{aligned}$$

Here (a) is by Bayes' rule and (b) is by the triangle inequality.

We will now bound the quantity C^+ in a way that does not depend on the specific assignment of $\pm$ to X_i, so the same bound will hold symmetrically for C^-. Using the identity $ab - \widehat{a}\widehat{b} = ab - a\widehat{b} + a\widehat{b} - \widehat{a}\widehat{b} = a(b-\widehat{b}) + \widehat{b}(a-\widehat{a})$, the triangle inequality, and the definition of $\lambda, \widehat{\lambda}$, we have

$$
\begin{aligned}
&C^+ \\
&= \sum_{x_{\mathcal{S}}} \Big| \lambda_i(x_{\mathcal{S}}) \frac{\mathsf{P}(X_u = +, X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}})}{\mathsf{P}(X_i = + | X_{\mathcal{S}} = x_{\mathcal{S}})} \\
&\qquad - \widehat{\lambda}_i(x_{\mathcal{S}}) \frac{\widehat{\mathsf{P}}(X_u = +, X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}})}{\widehat{\mathsf{P}}(X_i = + | X_{\mathcal{S}} = x_{\mathcal{S}})} \Big| \\
&\le \sum_{x_{\mathcal{S}}} \Big| \mathsf{P}(X_u = +, X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) \\
&\qquad \times \Big(\frac{\lambda_i(x_{\mathcal{S}})}{\mathsf{P}(X_i = + | X_{\mathcal{S}} = x_{\mathcal{S}})} - \frac{\widehat{\lambda}_i(x_{\mathcal{S}})}{\widehat{\mathsf{P}}(X_i = + | X_{\mathcal{S}} = x_{\mathcal{S}})} \Big) \Big| \\
&\quad + \sum_{x_{\mathcal{S}}} \Big| \frac{\widehat{\lambda}_i(x_{\mathcal{S}})}{\widehat{\mathsf{P}}(X_i = + | X_{\mathcal{S}} = x_{\mathcal{S}})} \Big(\mathsf{P}(X_u = +, X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) \\
&\qquad - \widehat{\mathsf{P}}(X_u = +, X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) \Big) \Big| \\
&= 2 \sum_{x_{\mathcal{S}}} \mathsf{P}(X_u = +, X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) \\
&\qquad \times \Big| \mathsf{P}(X_i = - | X_{\mathcal{S}} = x_{\mathcal{S}}) - \widehat{\mathsf{P}}(X_i = - | X_{\mathcal{S}} = x_{\mathcal{S}}) \Big| \\
&\quad + 2 \sum_{x_{\mathcal{S}}} \widehat{\mathsf{P}}(X_i = - | X_{\mathcal{S}} = x_{\mathcal{S}}) \Big| \mathsf{P}(X_u = +, X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) \\
&\qquad - \widehat{\mathsf{P}}(X_u = +, X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) \Big| \qquad (19)
\end{aligned}
$$

The latter sum is bounded by $2 \cdot 2^{|\mathcal{S}|}\gamma$. In order to bound the first sum in (19) we write

$$
\begin{aligned}
&\Big| \mathsf{P}(X_i = - | X_{\mathcal{S}} = x_{\mathcal{S}}) - \widehat{\mathsf{P}}(X_i = - | X_{\mathcal{S}} = x_{\mathcal{S}}) \Big| \\
&= \Big| \frac{\mathsf{P}(X_i = -, X_{\mathcal{S}} = x_{\mathcal{S}})}{\mathsf{P}(X_{\mathcal{S}} = x_{\mathcal{S}})} - \frac{\widehat{\mathsf{P}}(X_i = -, X_{\mathcal{S}} = x_{\mathcal{S}})}{\widehat{\mathsf{P}}(X_{\mathcal{S}} = x_{\mathcal{S}})} \Big| \\
&\le \Big| \frac{\mathsf{P}(X_i = -, X_{\mathcal{S}} = x_{\mathcal{S}})}{\mathsf{P}(X_{\mathcal{S}} = x_{\mathcal{S}})} - \frac{\widehat{\mathsf{P}}(X_i = -, X_{\mathcal{S}} = x_{\mathcal{S}})}{\mathsf{P}(X_{\mathcal{S}} = x_{\mathcal{S}})} \Big| \\
&\qquad + \Big| \frac{\widehat{\mathsf{P}}(X_i = -, X_{\mathcal{S}} = x_{\mathcal{S}})}{\mathsf{P}(X_{\mathcal{S}} = x_{\mathcal{S}})} - \frac{\widehat{\mathsf{P}}(X_i = -, X_{\mathcal{S}} = x_{\mathcal{S}})}{\widehat{\mathsf{P}}(X_{\mathcal{S}} = x_{\mathcal{S}})} \Big| \\
&\le \frac{2\gamma}{q} ,
\end{aligned}
$$

where $q := \delta^\ell \le \delta^{|\mathcal{S}|} \le \min_{x_{\mathcal{S}}} \mathsf{P}(X_{\mathcal{S}} = x_{\mathcal{S}})$. Plugging this into (19), the sum over $x_{\mathcal{S}}$ marginalizes over these variables and we obtain

$$
C^+ \le \frac{4\gamma}{q} \cdot \mathsf{P}(X_u = +, X_i = +) + 2^{|\mathcal{S}|+1}\gamma \le \frac{6\gamma}{q} .
$$

Here we used the fact that $q^{-1} = \delta^{-\ell} \ge 2^{|\mathcal{S}|}$, since $\delta \le 1/2$.

The same bound holds for C^-, so

$$
\begin{aligned}
&\big| \nu^{\mathtt{avg}}_{u|i;S} - \widehat{\nu}^{\mathtt{avg}}_{u|i;S} \big| \\
&= \big| \mathsf{E}_{X_{\mathcal{S}} \sim \mathsf{P}} \big(\lambda_i(X_{\mathcal{S}}) | \nu_{u|i;X_{\mathcal{S}}} | \big) - \mathsf{E}_{X_{\mathcal{S}} \sim \widehat{\mathsf{P}}} \big(\widehat{\lambda}_i(X_{\mathcal{S}}) | \widehat{\nu}_{u|i;X_{\mathcal{S}}} | \big) \big| \\
&\le \frac{12\gamma}{q} .
\end{aligned}
$$

Choosing $\gamma = \epsilon\delta^\ell/12$, we get the desired accuracy and our earlier choice of n evaluates to

$$
n = (2\gamma^2)^{-1} \log \left(\frac{2(\ell+2)(2p)^{\ell+2}}{\zeta} \right) \le \frac{144(\ell+3)}{\epsilon^2 \delta^{2\ell}} \log \frac{p}{\zeta} . \qquad \square
$$