Which graphical models are difficult to learn?

José Bento
Department of Electrical Engineering
Stanford University
jbento@stanford.edu

Andrea Montanari
Department of Electrical Engineering and
Department of Statistics
Stanford University
montanari@stanford.edu

Abstract

We consider the problem of learning the structure of Ising models (pairwise binary Markov random fields) from i.i.d. samples. While several methods have been proposed to accomplish this task, their relative merits and limitations remain somewhat obscure. By analyzing a number of concrete examples, we show that low-complexity algorithms systematically fail when the Markov random field develops long-range correlations. More precisely, this phenomenon appears to be related to the Ising model phase transition (although it does not coincide with it).

1   Introduction and main results

Given a graph $G = (V = [p], E)$, and a positive parameter $\theta > 0$ the *ferromagnetic Ising model on* G is the pairwise Markov random field

$$\mu_{G,\theta}(\underline{x}) = \frac{1}{Z_{G,\theta}} \prod_{(i,j) \in E} e^{\theta x_i x_j} \tag{1}$$

over binary variables $\underline{x} = (x_1, x_2, \ldots, x_p)$. Apart from being one of the most studied models in statistical mechanics, the Ising model is a prototypical undirected graphical model, with applications in computer vision, clustering and spatial statistics. Its obvious generalization to edge-dependent parameters θ_{ij}, $(i,j) \in E$ is of interest as well, and will be introduced in Section 1.2.2. (Let us stress that we follow the statistical mechanics convention of calling (1) an Ising model for any graph G.)

In this paper we study the following structural learning problem: *Given n i.i.d. samples $\underline{x}^{(1)}$, $\underline{x}^{(2)}, \ldots, \underline{x}^{(n)}$ with distribution $\mu_{G,\theta}(\,\cdot\,)$, reconstruct the graph G.* For the sake of simplicity, we assume that the parameter θ is known, and that G has no double edges (it is a 'simple' graph).

The graph learning problem is solvable with unbounded sample complexity, and computational resources [1]. The question we address is: for which classes of graphs and values of the parameter θ is the problem solvable under appropriate complexity constraints? More precisely, given an algorithm Alg, a graph G, a value θ of the model parameter, and a small $\delta > 0$, the sample complexity is defined as

$$n_{\mathsf{Alg}}(G,\theta) \equiv \inf \left\{ n \in \mathbb{N} : \mathbb{P}_{n,G,\theta}\{\mathsf{Alg}(\underline{x}^{(1)}, \ldots, \underline{x}^{(n)}) = G\} \geq 1 - \delta \right\}, \tag{2}$$

where $\mathbb{P}_{n,G,\theta}$ denotes probability with respect to n i.i.d. samples with distribution $\mu_{G,\theta}$. Further, we let $\chi_{\mathsf{Alg}}(G,\theta)$ denote the number of operations of the algorithm Alg, when run on $n_{\mathsf{Alg}}(G,\theta)$ samples.[1]

---

[1]For the algorithms analyzed in this paper, the behavior of n_{Alg} and χ_{Alg} does not change significantly if we require only 'approximate' reconstruction (e.g. in graph distance).

1

The general problem is therefore to characterize the functions $n_{\mathsf{Alg}}(G,\theta)$ and $\chi_{\mathsf{Alg}}(G,\theta)$, in particular for an optimal choice of the algorithm. General bounds on $n_{\mathsf{Alg}}(G,\theta)$ have been given in [2, 3], under the assumption of unbounded computational resources. A general charactrization of how well low complexity algorithms can perform is therefore lacking. Although we cannot prove such a general characterization, in this paper we estimate n_{Alg} and χ_{Alg} for a number of graph models, as a function of θ, and unveil a fascinating universal pattern: *when the model (1) develops long range correlations, low-complexity algorithms fail.* Under the Ising model, the variables $\{x_i\}_{i \in V}$ become strongly correlated for θ large. For a large class of graphs with degree bounded by Δ, this phenomenon corresponds to a phase transition beyond some critical value of θ uniformly bounded in p, with typically $\theta_{\mathrm{crit}} \leq \mathrm{const.}/\Delta$. In the examples discussed below, the failure of low-complexity algorithms appears to be related to this phase transition (although it does not coincide with it).

1.1 A toy example: the thresholding algorithm

In order to illustrate the interplay between graph structure, sample complexity and interaction strength θ, it is instructive to consider a warmup example. The thresholding algorithm reconstructs G by thresholding the empirical correlations

$$\widehat{C}_{ij} \equiv \frac{1}{n} \sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)} \qquad \text{for } i,j \in V. \tag{3}$$

THRESHOLDING(samples $\{x^{(\ell)}\}$, threshold τ)

1:     Compute the empirical correlations $\{\widehat{C}_{ij}\}_{(i,j) \in V \times V}$;
2:     For each $(i,j) \in V \times V$
3:        If $\widehat{C}_{ij} \geq \tau$, set $(i,j) \in E$;

We will denote this algorithm by $\mathsf{Thr}(\tau)$. Notice that its complexity is dominated by the computation of the empirical correlations, i.e. $\chi_{\mathsf{Thr}(\tau)} = O(p^2 n)$. The sample complexity $n_{\mathsf{Thr}(\tau)}$ can be bounded for specific classes of graphs as follows (the proofs are straightforward and omitted from this paper).

Theorem 1.1. *If G has maximum degree $\Delta > 1$ and if $\theta < \mathrm{atanh}(1/(2\Delta))$ then there exists $\tau = \tau(\theta)$ such that*

$$n_{\mathsf{Thr}(\tau)}(G,\theta) \leq \frac{8}{(\tanh\theta - \frac{1}{2\Delta})^2} \ \log \frac{2p}{\delta} \ . \tag{4}$$

Further, the choice $\tau(\theta) = (\tanh\theta + (1/2\Delta))/2$ achieves this bound.

Theorem 1.2. *There exists a numerical constant K such that the following is true. If $\Delta > 3$ and $\theta > K/\Delta$, there are graphs of bounded degree Δ such that for any τ, $n_{\mathsf{Thr}(\tau)} = \infty$, i.e. the thresholding algorithm always fails with high probability.*

These results confirm the idea that the failure of low-complexity algorithms is related to long-range correlations in the underlying graphical model. If the graph G is a tree, then correlations between far apart variables x_i, x_j decay exponentially with the distance between vertices i, j. The same happens on bounded-degree graphs if $\theta \leq \mathrm{const.}/\Delta$. However, for $\theta > \mathrm{const.}/\Delta$, there exists families of bounded degree graphs with long-range correlations.

1.2 More sophisticated algorithms

In this section we characterize $\chi_{\mathsf{Alg}}(G,\theta)$ and $n_{\mathsf{Alg}}(G,\theta)$ for more advanced algorithms. We again obtain very distinct behaviors of these algorithms depending on long range correlations. Due to space limitations, we focus on two type of algorithms and only outline the proof of our most challenging result, namely Theorem 1.6.

In the following we denote by ∂i the neighborhood of a node $i \in G$ ($i \notin \partial i$), and assume the degree to be bounded: $|\partial i| \leq \Delta$.

1.2.1 Local Independence Test

A recurring approach to structural learning consists in exploiting the conditional independence structure encoded by the graph [1, 4, 5, 6].

2

Let us consider, to be definite, the approach of [4], specializing it to the model (1). Fix a vertex r, whose neighborhood we want to reconstruct, and consider the conditional distribution of x_r given its neighbors[2]: $\mu_{G,\theta}(x_r|\underline{x}_{\partial r})$. Any change of x_i, $i \in \partial r$, produces a change in this distribution which is bounded away from 0. Let U be a candidate neighborhood, and assume $U \subseteq \partial r$. Then changing the value of x_j, $j \in U$ will produce a noticeable change in the marginal of X_r, even if we condition on the remaining values in U and in any W, $|W| \leq \Delta$. On the other hand, if $U \nsubseteq \partial r$, then it is possible to find W (with $|W| \leq \Delta$) and a node $i \in U$ such that, changing its value after fixing all other values in $U \cup W$ will produce no noticeable change in the conditional marginal. (Just choose $i \in U \backslash \partial r$ and $W = \partial r \backslash U$). This procedure allows us to distinguish subsets of ∂r from other sets of vertices, thus motivating the following algorithm.

LOCAL INDEPENDENCE TEST(samples $\{x^{(\ell)}\}$, thresholds (ϵ, γ))
1: Select a node $r \in V$;
2: Set as its neighborhood the largest candidate neighbor U of
    size at most Δ for which the score function $\mathrm{SCORE}(U) > \epsilon/2$;
3: Repeat for all nodes $r \in V$;

The score function $\mathrm{SCORE}(\cdot)$ depends on $(\{x^{(\ell)}\}, \Delta, \gamma)$ and is defined as follows,

$$\min_{W,j} \max_{x_i, \underline{x}_W, \underline{x}_U, x_j} |\widehat{\mathbb{P}}_{n,G,\theta}\{X_i = x_i | \underline{X}_W = \underline{x}_W, \underline{X}_U = \underline{x}_U\} -$$
$$\widehat{\mathbb{P}}_{n,G,\theta}\{X_i = x_i | \underline{X}_W = \underline{x}_W, \underline{X}_{U \backslash j} = \underline{x}_{U \backslash j}, X_j = x_j\}|. \quad (5)$$

In the minimum, $|W| \leq \Delta$ and $j \in U$. In the maximum, the values must be such that

$$\widehat{\mathbb{P}}_{n,G,\theta}\{\underline{X}_W = \underline{x}_W, \underline{X}_U = \underline{x}_U\} > \gamma/2, \quad \widehat{\mathbb{P}}_{n,G,\theta}\{\underline{X}_W = \underline{x}_W, \underline{X}_{U \backslash j} = \underline{x}_{U \backslash j}, X_j = x_j\} > \gamma/2$$

$\widehat{\mathbb{P}}_{n,G,\theta}$ is the empirical distribution calculated from the samples $\{x^{(\ell)}\}$. We denote this algorithm by $\mathsf{Ind}(\epsilon, \gamma)$. The search over candidate neighbors U, the search for minima and maxima in the computation of the $\mathrm{SCORE}(U)$ and the computation of $\widehat{\mathbb{P}}_{n,G,\theta}$ all contribute for $\chi_{\mathsf{Ind}}(G, \theta)$.

Both theorems that follow are consequences of the analysis of [4].

Theorem 1.3. *Let G be a graph of bounded degree $\Delta \geq 1$. For every θ there exists (ϵ, γ), and a numerical constant K, such that*

$$n_{\mathsf{Ind}(\epsilon,\gamma)}(G, \theta) \leq \frac{100\Delta}{\epsilon^2 \gamma^4} \log \frac{2p}{\delta}, \quad \chi_{\mathsf{Ind}(\epsilon,\gamma)}(G, \theta) \leq K (2p)^{2\Delta+1} \log p.$$

More specifically, one can take $\epsilon = \frac{1}{4}\sinh(2\theta)$, $\gamma = e^{-4\Delta\theta} 2^{-2\Delta}$.

This first result implies in particular that G can be reconstructed with polynomial complexity for any bounded Δ. However, the degree of such polynomial is pretty high and non-uniform in Δ. This makes the above approach impractical.

A way out was proposed in [4]. The idea is to identify a set of 'potential neighbors' of vertex r via thresholding:

$$B(r) = \{i \in V : \widehat{C}_{ri} > \kappa/2\}, \quad (6)$$

For each node $r \in V$, we evaluate $\mathrm{SCORE}(U)$ by restricting the minimum in Eq. (5) over $W \subseteq B(r)$, and search only over $U \subseteq B(r)$. We call this algorithm $\mathsf{IndD}(\epsilon, \gamma, \kappa)$. The basic intuition here is that C_{ri} decreases rapidly with the graph distance between vertices r and i. As mentioned above, this is true at small θ.

Theorem 1.4. *Let G be a graph of bounded degree $\Delta \geq 1$. Assume that $\theta < K/\Delta$ for some small enough constant K. Then there exists ϵ, γ, κ such that*

$$n_{\mathsf{IndD}(\epsilon,\gamma,\kappa)}(G, \theta) \leq 8(\kappa^2 + 8^\Delta) \log \frac{4p}{\delta}, \quad \chi_{\mathsf{IndD}(\epsilon,\gamma,\kappa)}(G, \theta) \leq K' p \Delta^{\Delta \frac{\log(4/\kappa)}{\alpha}} + K' \Delta p^2 \log p.$$

More specifically, we can take $\kappa = \tanh\theta$, $\epsilon = \frac{1}{4}\sinh(2\theta)$ and $\gamma = e^{-4\Delta\theta} 2^{-2\Delta}$.

---

[2]If $\underline{a}$ is a vector and R is a set of indices then we denote by $\underline{a}_R$ the vector formed by the components of $\underline{a}$ with index in R.

3

1.2.2 Regularized Pseudo-Likelihoods

A different approach to the learning problem consists in maximizing an appropriate empirical likelihood function [7, 8, 9, 10, 13]. To control the fluctuations caused by the limited number of samples, and select sparse graphs a regularization term is often added [7, 8, 9, 10, 11, 12, 13].

As a specific low complexity implementation of this idea, we consider the ℓ_1-regularized pseudo-likelihood method of [7]. For each node r, the following likelihood function is considered

$$L(\underline{\theta}; \{x^{(\ell)}\}) = -\frac{1}{n} \sum_{\ell=1}^{n} \log \mathbb{P}_{n,G,\underline{\theta}}(x_r^{(\ell)}|x_{\backslash r}^{(\ell)}) \tag{7}$$

where $\underline{x}_{\backslash r} = \underline{x}_{V\backslash r} = \{x_i : i \in V \backslash r\}$ is the vector of all variables except x_r and $\mathbb{P}_{n,G,\underline{\theta}}$ is defined from the following extension of (1),

$$\mu_{G,\underline{\theta}}(\underline{x}) = \frac{1}{Z_{G,\underline{\theta}}} \prod_{i,j \in V} e^{\theta_{ij} x_i x_j} \tag{8}$$

where $\underline{\theta} = \{\theta_{ij}\}_{i,j \in V}$ is a vector of real parameters. Model (1) corresponds to $\theta_{ij} = 0,\ \forall (i,j) \notin E$ and $\theta_{ij} = \theta,\ \forall (i,j) \in E$.

The function $L(\underline{\theta}; \{x^{(\ell)}\})$ depends only on $\underline{\theta}_{r,\cdot} = \{\theta_{rj}, j \in \partial r\}$ and is used to estimate the neighborhood of each node by the following algorithm, $\mathsf{Rlr}(\lambda)$,

---
REGULARIZED LOGISTIC REGRESSION(samples $\{x^{(\ell)}\}$, regularization (λ))
1:     Select a node $r \in V$;
2:     Calculate $\hat{\underline{\theta}}_{r,\cdot} = \arg\min_{\underline{\theta}_{r,\cdot} \in \mathbb{R}^{p-1}} \{L(\underline{\theta}_{r,\cdot}; \{x^{(\ell)}\}) + \lambda ||\underline{\theta}_{r,\cdot}||_1\}$;
3:      If $\hat{\theta}_{rj} > 0$, set $(r,j) \in E$;

Our first result shows that $\mathsf{Rlr}(\lambda)$ indeed reconstructs G if θ is sufficiently small.

Theorem 1.5. *There exists numerical constants K_1, K_2, K_3, such that the following is true. Let G be a graph with degree bounded by $\Delta \geq 3$. If $\theta \leq K_1/\Delta$, then there exist λ such that*

$$n_{\mathsf{Rlr}(\lambda)}(G,\theta) \leq K_2\, \theta^{-2}\, \Delta\, \log \frac{8p^2}{\delta}\,. \tag{9}$$

Further, the above holds with $\lambda = K_3\, \theta\, \Delta^{-1/2}$.

This theorem is proved by noting that for $\theta \leq K_1/\Delta$ correlations decay exponentially, which makes all conditions in Theorem 1 of [7] (denoted there by A1 and A2) hold, and then computing the probability of success as a function of n, while strenghtening the error bounds of [7].

In order to prove a converse to the above result, we need to make some assumptions on λ. Given $\theta > 0$, we say that λ is 'reasonable for that value of θ if the following conditions old: (i) $\mathsf{Rlr}(\lambda)$ is successful with probability larger than $1/2$ on any star graph (a graph composed by a vertex r connected to Δ neighbors, plus isolated vertices); (ii) $\lambda \leq \delta(n)$ for some sequence $\delta(n) \downarrow 0$.

Theorem 1.6. *There exists a numerical constant K such that the following happens. If $\Delta > 3$, $\theta > K/\Delta$, then there exists graphs G of degree bounded by Δ such that for all reasonable λ, $n_{\mathsf{Rlr}(\lambda)}(G) = \infty$, i.e. regularized logistic regression fails with high probability.*

The graphs for which regularized logistic regression fails are not contrived examples. Indeed we will prove that the claim in the last theorem holds with high probability when G is a uniformly random graph of regular degree Δ.

The proof Theorem 1.6 is based on showing that an appropriate *incoherence condition* is necessary for Rlr to successfully reconstruct G. The analogous result was proven in [14] for model selection using the Lasso. In this paper we show that such a condition is also necessary when the underlying model is an Ising model. Notice that, given the graph G, checking the incoherence condition is NP-hard for general (non-ferromagnetic) Ising model, and requires significant computational effort

4
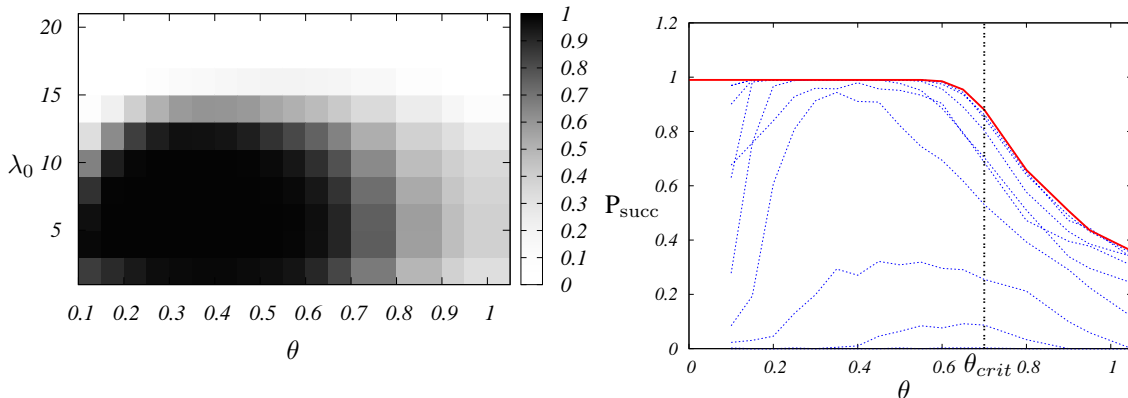
Figure 1: Learning random subgraphs of a 7×7 ($p = 49$) two-dimensional grid from $n = 4500$ Ising models samples, using regularized logistic regression. Left: success probability as a function of the model parameter θ and of the regularization parameter λ_0 (darker corresponds to highest probability). Right: the same data plotted for several choices of λ versus θ. The vertical line corresponds to the model critical temperature. The thick line is an envelope of the curves obtained for different λ, and should correspond to optimal regularization.

even in the ferromagnetic case. Hence the incoherence condition does not provide, by itself, a clear picture of which graph structure are difficult to learn. We will instead show how to evaluate it on specific graph families.

Under the restriction $\lambda \to 0$ the solutions given by Rlr converge to $\underline{\theta}^*$ with n [7]. Thus, for large n we can expand L around $\underline{\theta}^*$ to second order in $(\underline{\theta} - \underline{\theta}^*)$. When we add the regularization term to L we obtain a quadratic model analogous the Lasso plus the error term due to the quadratic approximation. It is thus not surprising that, when $\lambda \to 0$ the incoherence condition introduced for the Lasso in [14] is also relevant for the Ising model.

2 Numerical experiments

In order to explore the practical relevance of the above results, we carried out extensive numerical simulations using the regularized logistic regression algorithm Rlr(λ). Among other learning algorithms, Rlr(λ) strikes a good balance of complexity and performance. Samples from the Ising model (1) where generated using Gibbs sampling (a.k.a. Glauber dynamics). Mixing time can be very large for $\theta \geq \theta_{\mathrm{crit}}$, and was estimated using the time required for the overall bias to change sign (this is a quite conservative estimate at low temperature). Generating the samples $\{\underline{x}^{(\ell)}\}$ was indeed the bulk of our computational effort and took about 50 days CPU time on Pentium Dual Core processors (we show here only part of these data). Notice that Rlr(λ) had been tested in [7] only on tree graphs G, or in the weakly coupled regime $\theta < \theta_{\mathrm{crit}}$. In these cases sampling from the Ising model is easy, but structural learning is also intrinsically easier.

Figure reports the success probability of Rlr(λ) when applied to random subgraphs of a 7×7 two-dimensional grid. Each such graphs was obtained by removing each edge independently with probability $\rho = 0.3$. Success probability was estimated by applying Rlr(λ) to each vertex of 8 graphs (thus averaging over 392 runs of Rlr(λ)), using $n = 4500$ samples. We scaled the regularization parameter as $\lambda = 2\lambda_0\theta(\log p/n)^{1/2}$ (this choice is motivated by the algorithm analysis and is empirically the most satisfactory), and searched over λ_0.

The data clearly illustrate the phenomenon discussed. Despite the large number of samples $n \gg \log p$, when θ crosses a threshold, the algorithm starts performing poorly irrespective of λ. Intriguingly, this threshold is not far from the critical point of the Ising model on a randomly diluted grid $\theta_{\mathrm{crit}}(\rho = 0.3) \approx 0.7$ [15, 16].

5
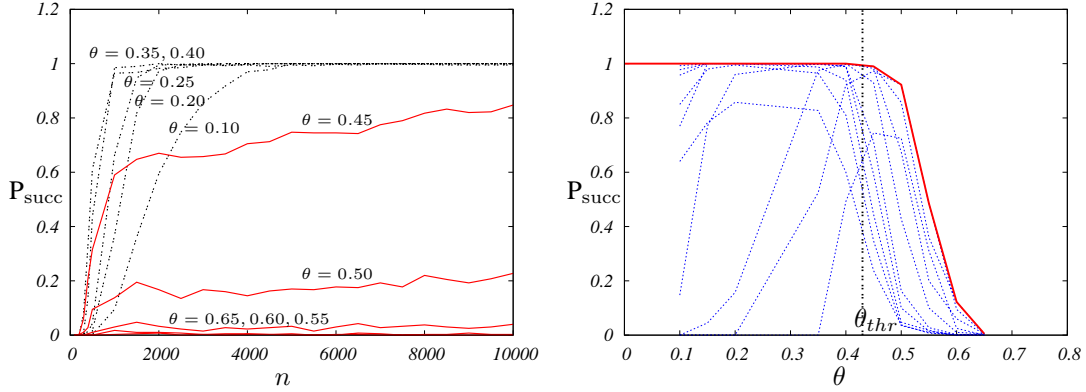
Figure 2: Learning uniformly random graphs of degree 4 from Ising models samples, using Rlr. Left: success probability as a function of the number of samples n for several values of θ. Right: the same data plotted for several choices of λ versus θ as in Fig. 1, right panel.

Figure 2 presents similar data when G is a uniformly random graph of degree $\Delta = 4$, over $p = 50$ vertices. The evolution of the success probability with n clearly shows a dichotomy. When θ is below a threshold, a small number of samples is sufficient to reconstruct G with high probability. Above the threshold even $n = 10^4$ samples are to few. In this case we can predict the threshold analytically, cf. Lemma 3.3 below, and get $\theta_{\text{thr}}(\Delta = 4) \approx 0.4203$, which compares favorably with the data.

3 Proofs

In order to prove Theorem 1.6, we need a few auxiliary results. It is convenient to introduce some notations. If M is a matrix and R, P are index sets then $M_{R\,P}$ denotes the submatrix with row indices in R and column indices in P. As above, we let r be the vertex whose neighborhood we are trying to reconstruct and define $S = \partial r$, $S^c = V \setminus \partial r \cup r$. Since the cost function $L(\underline{\theta}; \{x^{(\ell)}\}) + \lambda ||\underline{\theta}||_1$ only depend on $\underline{\theta}$ through its components $\underline{\theta}_{r,\cdot} = \{\theta_{rj}\}$, we will hereafter neglect all the other parameters and write $\underline{\theta}$ as a shorthand of $\underline{\theta}_{r,\cdot}$.

Let $\hat{z}^*$ be a subgradient of $||\underline{\theta}||_1$ evaluated at the true parameters values, $\underline{\theta}^* = \{\theta_{rj} : \theta_{ij} = 0, \ \forall j \notin \partial r, \theta_{rj} = \theta, \ \forall j \in \partial r\}$. Let $\hat{\underline{\theta}}^n$ be the parameter estimate returned by $\text{Rlr}(\lambda)$ when the number of samples is n. Note that, since we assumed $\underline{\theta}^* \geq 0$, $\hat{z}^*_S = \mathbb{1}$. Define $Q^n(\underline{\theta}, ; \{\underline{x}^{(\ell)}\})$ to be the Hessian of $L(\underline{\theta}; \{x^{(\ell)}\})$ and $Q(\underline{\theta}) = \lim_{n \to \infty} Q^n(\underline{\theta}, ; \{\underline{x}^{(\ell)}\})$. By the law of large numbers $Q(\underline{\theta})$ is the Hessian of $\mathbb{E}_{G,\theta} \log \mathbb{P}_{G,\underline{\theta}}(X_r|X_{\backslash r})$ where $\mathbb{E}_{G,\theta}$ is the expectation with respect to (8) and $\underline{X}$ is a random variable distributed according to (8). We will denote the maximum and minimum eigenvalue of a symmetric matrix M by $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ respectively.

We will omit arguments whenever clear from the context. Any quantity evaluated at the true parameter values will be represented with a $*$, e.g. $Q^* = Q(\theta^*)$. Quantities under a $\wedge$ depend on n. Throughout this section G is a graph of maximum degree Δ.

3.1 Proof of Theorem 1.6

Our first auxiliary results establishes that, if λ is small, then $||Q^*_{S^c S} Q^{*-1}_{SS} \hat{z}^*_S||_\infty > 1$ is a sufficient condition for the failure of $\text{Rlr}(\lambda)$.

Lemma 3.1. *Assume* $[Q^*_{S^c S} Q^{*-1}_{SS} \hat{z}^*_S]_i \geq 1 + \epsilon$ *for some* $\epsilon > 0$ *and some row* $i \in V$, $\sigma_{\min}(Q^*_{SS}) \geq C_{\min} > 0$, *and* $\lambda < \sqrt{C^3_{\min}\epsilon/2^9\Delta^4}$. *Then the success probability of* $\text{Rlr}(\lambda)$ *is upper bounded as*

$$\text{P}_{\text{succ}} \leq 4\Delta^2 e^{-n\delta_A^2} + 2\Delta\, e^{-n\lambda^2\delta_B^2} \tag{10}$$

where $\delta_A = (C^2_{\min}/100\Delta^2)\epsilon$ *and* $\delta_B = (C_{\min}/8\Delta)\epsilon$.

6

The next Lemma implies that, for λ to be 'reasonable' (in the sense introduced in Section 1.2.2), $n\lambda^2$ must be unbounded.

Lemma 3.2. *There exist $M = M(K, \theta) > 0$ for $\theta > 0$ such that the following is true: If G is the graph with only one edge between nodes r and i and $n\lambda^2 \leq K$, then*

$$\mathrm{P}_{\mathrm{succ}} \leq e^{-M(K,\theta)p} + e^{-n(1-\tanh\theta)^2/32} . \tag{11}$$

Finally, our key result shows that the condition $||Q^*_{S^cS}Q^*_{SS}{}^{-1}\hat{z}^*_S||_\infty \leq 1$ is violated with high probability for large random graphs. The proof of this result relies on a local weak convergence result for ferromagnetic Ising models on random graphs proved in [17].

Lemma 3.3. *Let G be a uniformly random regular graph of degree $\Delta > 3$, and $\epsilon > 0$ be sufficiently small. Then, there exists $\theta_{\mathrm{thr}}(\Delta, \epsilon)$ such that, for $\theta > \theta_{\mathrm{thr}}(\Delta, \epsilon)$, $||Q^*_{S^cS}Q^*_{SS}{}^{-1}\hat{z}^*_S||_\infty \geq 1 + \epsilon$ with probability converging to 1 as $p \to \infty$.*

Furthermore, for large Δ, $\theta_{\mathrm{thr}}(\Delta, 0+) = \tilde{\theta}\Delta^{-1}(1 + o(1))$. The constant $\tilde{\theta}$ is given by $\tilde{\theta} = \tanh\bar{h})/\bar{h}$ and $\bar{h}$ is the unique positive solution of $\bar{h}\tanh\bar{h} = (1 - \tanh^2\bar{h})^2$. Finally, there exist $C_{\min} > 0$ dependent only on Δ and θ such that $\sigma_{\min}(Q^*_{SS}) \geq C_{\min}$ with probability converging to 1 as $p \to \infty$.

The proofs of Lemmas 3.1 and 3.3 are sketched in the next subsection. Lemma 3.2 is more straightforward and we omit its proof for space reasons.

Proof. (Theorem 1.6) Fix $\Delta > 3$, $\theta > K/\Delta$ (where K is a large enough constant independent of Δ), and $\epsilon, C_{\min} > 0$ and both small enough. By Lemma 3.3, for any p large enough we can choose a Δ-regular graph $G_p = (V = [p], E_p)$ and a vertex $r \in V$ such that $|Q^*_{S^cS}Q^*_{SS}{}^{-1}\mathbb{1}_S|_i > 1 + \epsilon$ for some $i \in V \setminus r$.

By Theorem 1 in [4] we can assume, without loss of generality $n > K'\Delta\log p$ for some small constant K'. Further by Lemma 3.2, $n\lambda^2 \geq F(p)$ for some $F(p) \uparrow \infty$ as $p \to \infty$ and the condition of Lemma 3.1 on λ is satisfied since by the "reasonable" assumption $\lambda \to 0$ with n. Using these results in Eq. (10) of Lemma 3.1 we get the following upper bound on the success probability

$$\mathrm{P}_{\mathrm{succ}}(G_p) \leq 4\Delta^2 p^{-\delta_A^2 K'\Delta} + 2\Delta e^{-nF(p)\delta_B^2} . \tag{12}$$

In particular $\mathrm{P}_{\mathrm{succ}}(G_p) \to 0$ as $p \to \infty$. $\qquad\qquad\square$

3.2 Proofs of auxiliary lemmas

Proof. (Lemma 3.1) We will show that under the assumptions of the lemma and if $\hat{\underline{\theta}} = (\hat{\underline{\theta}}_S, \hat{\underline{\theta}}_{S^C}) = (\hat{\underline{\theta}}_S, 0)$ then the probability that the i component of any subgradient of $L(\underline{\theta}; \{x^{(\ell)}\}) + \lambda||\underline{\theta}||_1$ vanishes for any $\hat{\underline{\theta}}_S > 0$ (component wise) is upper bounded as in Eq. (10). To simplify notation we will omit $\{x^{(\ell)}\}$ in all the expression derived from L.

Let $\hat{z}$ be a subgradient of $||\underline{\theta}||$ at $\hat{\underline{\theta}}$ and assume $\nabla L(\hat{\underline{\theta}}) + \lambda\hat{z} = 0$. An application of the mean value theorem yields

$$\nabla^2 L(\underline{\theta}^*)[\hat{\underline{\theta}} - \underline{\theta}^*] = W^n - \lambda\hat{z} + R^n , \tag{13}$$

where $W^n = -\nabla L(\underline{\theta}^*)$ and $[R^n]_j = [\nabla^2 L(\bar{\underline{\theta}}^{(j)}) - \nabla^2 L(\underline{\theta}^*)]_j^T(\hat{\underline{\theta}} - \underline{\theta}^*)$ with $\bar{\underline{\theta}}^{(j)}$ a point in the line from $\hat{\underline{\theta}}$ to $\underline{\theta}^*$. Notice that by definition $\nabla^2 L(\underline{\theta}^*) = Q^{n*} = Q^n(\underline{\theta}^*)$. To simplify notation we will omit the $*$ in all Q^{n*}. All Q^n in this proof are thus evaluated at $\underline{\theta}^*$.

Breaking this expression into its S and S^c components and since $\hat{\underline{\theta}}_{S^C} = \underline{\theta}^*_{S^C} = 0$ we can eliminate $\hat{\underline{\theta}}_S - \underline{\theta}^*_S$ from the two expressions obtained and write

$$[W^n_{S^C} - R^n_{S^C}] - Q^n_{S^CS}(Q^n_{SS})^{-1}[W^n_S - R^n_S] + \lambda Q^n_{S^CS}(Q^n_{SS})^{-1}\hat{z}_S = \lambda\hat{z}_{S^C} . \tag{14}$$

Now notice that $Q^n_{S^CS}(Q^n_{SS})^{-1} = T_1 + T_2 + T_3 + T_4$ where

$$T_1 = Q^*_{S^CS}[(Q^n_{SS})^{-1} - (Q^*_{SS})^{-1}] , \qquad\qquad T_2 = [Q^n_{S^CS} - Q^*_{S^CS}]Q^*_{SS}{}^{-1} ,$$
$$T_3 = [Q^n_{S^CS} - Q^*_{S^CS}][(Q^n_{SS})^{-1} - (Q^*_{SS})^{-1}] , \qquad T_4 = Q^*_{S^CS}Q^*_{SS}{}^{-1} .$$

7

We will assume that the samples $\{x^{(\ell)}\}$ are such that the following event holds

$$\mathcal{E} \equiv \{||Q_{SS}^n - Q_{SS}^*||_\infty < \xi_A, ||Q_{S^C S}^n - Q_{S^C S}^*||_\infty < \xi_B, ||W_S^n/\lambda||_\infty < \xi_C\}, \qquad (15)$$

where $\xi_A \equiv C_{\min}^2 \epsilon/(16\Delta)$, $\xi_B \equiv C_{\min}\epsilon/(8\sqrt{\Delta})$ and $\xi_C \equiv C_{\min}\epsilon/(8\Delta)$. Since $\mathbb{E}_{G,\theta}(Q^n) = Q^*$ and $\mathbb{E}_{G,\theta}(W^n) = 0$ and noticing that both Q^n and W^n are sums of bounded i.i.d. random variables, a simple application of Azuma-Hoeffding inequality upper bounds the probability of $\mathcal{E}$ as in (10).

From $\mathcal{E}$ it follows that $\sigma_{\min}(Q_{SS}^n) > \sigma_{\min}(Q_{SS}^*) - C_{\min}/2 > C_{\min}/2$. We can therefore lower bound the absolute value of the ith component of $\hat{z}_{S^C}$ by

$$|[Q_{S^C S}^* Q_{SS}^{*-1} \mathbb{1}_S]_i| - ||T_{1,i}||_\infty - ||T_{2,i}||_\infty - ||T_{3,i}||_\infty - \left|\frac{W_i^n}{\lambda}\right| - \left|\frac{R_i^n}{\lambda}\right| - \frac{\Delta}{C_{\min}}\left(\left\|\frac{W_S^n}{\lambda}\right\|_\infty + \left\|\frac{R_S^n}{\lambda}\right\|_\infty\right),$$

where the subscript i denotes the i-th row of a matrix.

The proof is completed by showing that the event $\mathcal{E}$ and the assumptions of the theorem imply that each of last 7 terms in this expression is smaller than $\epsilon/8$. Since $|[Q_{S^C S}^* Q_{SS}^{*-1}]_i^T \hat{z}_S^n| \geq 1 + \epsilon$ by assumption, this implies $|\hat{z}_i| \geq 1 + \epsilon/8 > 1$ which cannot be since any subgradient of the 1-norm has components of magnitude at most 1.

The last condition on $\mathcal{E}$ immediately bounds all terms involving W by $\epsilon/8$. Some straightforward manipulations imply (See Lemma 7 from [7])

$$||T_{1,i}||_\infty \leq \frac{\Delta}{C_{\min}^2}||Q_{SS}^n - Q_{SS}^*||_\infty, \qquad ||T_{2,i}||_\infty \leq \frac{\sqrt{\Delta}}{C_{\min}}||[Q_{S^C S}^n - Q_{S^C S}^*]_i||_\infty,$$

$$||T_{3,i}||_\infty \leq \frac{2\Delta}{C_{\min}^2}||Q_{SS}^n - Q_{SS}^*||_\infty ||[Q_{S^C S}^n - Q_{S^C S}^*]_i||_\infty,$$

and thus all will be bounded by $\epsilon/8$ when $\mathcal{E}$ holds. The upper bound of R^n follows along similar lines via an mean value theorem, and is deferred to a longer version of this paper. $\square$

Proof. (Lemma 3.3.) Let us state explicitly the local weak convergence result mentioned in Sec. 3.1. For $t \in \mathbb{N}$, let $\mathsf{T}(t) = (V_\mathsf{T}, E_\mathsf{T})$ be the regular rooted tree of t generations and define the associated Ising measure as

$$\mu_{\mathsf{T},\theta}^+(\underline{x}) = \frac{1}{Z_{\mathsf{T},\theta}} \prod_{(i,j)\in E_\mathsf{T}} e^{\theta x_i x_j} \prod_{i\in\partial\mathsf{T}(t)} e^{h^* x_i}. \qquad (16)$$

Here $\partial\mathsf{T}(t)$ is the set of leaves of $\mathsf{T}(t)$ and h^* is the unique positive solution of $h = (\Delta - 1)\,\mathrm{atanh}\,\{\tanh\theta\tanh h\}$. It can be proved using [17] and uniform continuity with respect to the 'external field' that non-trivial local expectations with respect to $\mu_{G,\theta}(\underline{x})$ converge to local expectations with respect to $\mu_{\mathsf{T},\theta}^+(\underline{x})$, as $p \to \infty$.

More precisely, let $\mathsf{B}_r(t)$ denote a ball of radius t around node $r \in G$ (the node whose neighborhood we are trying to reconstruct). For any fixed t, the probability that $\mathsf{B}_r(t)$ is not isomorphic to $\mathsf{T}(t)$ goes to 0 as $p \to \infty$. Let $g(\underline{x}_{\mathsf{B}_r(t)})$ be any function of the variables in $\mathsf{B}_r(t)$ such that $g(\underline{x}_{\mathsf{B}_r(t)}) = g(-\underline{x}_{\mathsf{B}_r(t)})$. Then almost surely over graph sequences G_p of uniformly random regular graphs with p nodes (expectations here are taken with respect to the measures (1) and (16))

$$\lim_{p\to\infty} \mathbb{E}_{G,\theta}\{g(\underline{X}_{\mathsf{B}_r(t)})\} = \mathbb{E}_{\mathsf{T}(t),\theta,+}\{g(\underline{X}_{\mathsf{T}(t)})\}. \qquad (17)$$

The proof consists in considering $[Q_{S^C S}^* Q_{SS}^{*-1} \hat{z}_S^*]_i$ for $t = \mathrm{dist}(r,i)$ finite. We then write $(Q_{SS}^*)_{lk} = \mathbb{E}\{g_{l,k}(\underline{X}_{\mathsf{B}_r(t)})\}$ and $(Q_{S^C S}^*)_{il} = \mathbb{E}\{g_{i,l}(\underline{X}_{\mathsf{B}_r(t)})\}$ for some functions $g_{\cdot,\cdot}(\underline{X}_{\mathsf{B}_r(t)})$ and apply the weak convergence result (17) to these expectations. We thus reduced the calculation of $[Q_{S^C S}^* Q_{SS}^{*-1} \hat{z}_S^*]_i$ to the calculation of expectations with respect to the tree measure (16). The latter can be implemented explicitly through a recursive procedure, with simplifications arising thanks to the tree symmetry and by taking $t \gg 1$. The actual calculations consist in a (very) long exercise in calculus and we omit them from this outline.

The lower bound on $\sigma_{\min}(Q_{SS}^*)$ is proved by a similar calculation. $\square$

Acknowledgments

This work was partially supported by a Terman fellowship, the NSF CAREER award CCF-0743978 and the NSF grant DMS-0806211 and by a Portuguese Doctoral FCT fellowship.

8

References

[1] P. Abbeel, D. Koller and A. Ng, "Learning factor graphs in polynomial time and sample complexity". Journal of Machine Learning Research., 2006, Vol. 7, 1743–1788.

[2] M. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting", arXiv:math/0702301v2 [math.ST], 2007.

[3] N. Santhanam, M. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions", arXiv:0905.2639v1 [cs.IT], 2009.

[4] G. Bresler, E. Mossel and A. Sly, "Reconstruction of Markov Random Fields from Samples: Some Observations and Algorithms",Proceedings of the 11th international workshop, APPROX 2008, and 12th international workshop RANDOM 2008, 2008 ,343–356.

[5] Csiszar and Z. Talata, "Consistent estimation of the basic neighborhood structure of Markov random fields", The Annals of Statistics, 2006, 34, Vol. 1, 123-145.

[6] N. Friedman, I. Nachman, and D. Peer, "Learning Bayesian network structure from massive datasets: The sparse candidate algorithm". In UAI, 1999.

[7] P. Ravikumar, M. Wainwright and J. Lafferty, "High-Dimensional Ising Model Selection Using l1-Regularized Logistic Regression", arXiv:0804.4202v1 [math.ST], 2008.

[8] M.Wainwright, P. Ravikumar, and J. Lafferty, "Inferring graphical model structure using l1-regularized pseudolikelihood", In NIPS, 2006.

[9] H. Höfling and R. Tibshirani, "Estimation of Sparse Binary Pairwise Markov Networks using Pseudo-likelihoods" , Journal of Machine Learning Research, 2009, Vol. 10, 883–906.

[10] O.Banerjee, L. El Ghaoui and A. d'Aspremont, "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data", Journal of Machine Learning Research, March 2008, Vol. 9, 485–516.

[11] M. Yuan and Y. Lin, "Model Selection and Estimation in Regression with Grouped Variables", J. Royal. Statist. Soc B, 2006, 68, Vol. 19,49–67.

[12] N. Meinshausen and P. Büuhlmann, "High dimensional graphs and variable selection with the lasso", Annals of Statistics, 2006, 34, Vol. 3.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society, Series B, 1994, Vol. 58, 267–288.

[14] P. Zhao, B. Yu, "On model selection consistency of Lasso", Journal of Machine. Learning Research 7, 25412563, 2006.

[15] D. Zobin, "Critical behavior of the bond-dilute two-dimensional Ising model", Phys. Rev., 1978 ,5, Vol. 18, 2387 – 2390.

[16] M. Fisher, "Critical Temperatures of Anisotropic Ising Lattices. II. General Upper Bounds", Phys. Rev. 162 ,Oct. 1967, Vol. 2, 480–485.

[17] A. Dembo and A. Montanari, "Ising Models on Locally Tree Like Graphs", Ann. Appl. Prob. (2008), to appear, arXiv:0804.4726v2 [math.PR]

9