arXiv:1110.1769v1 [stat.ML] 8 Oct 2011

On the trade-off between complexity and correlation decay in structural learning algorithms

José Bento* and Andrea Montanari†

October 11, 2011

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 often 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 $\underline{x} = (x_1, x_2, \dots, x_p)$, $x_i \in \{+1, -1\}$. Apart from being one of the best studied models in statistical mechanics [1, 2], the Ising model is a prototypical undirected graphical model. Since the seminal work of Hopfield [3] and Hinton and Sejnowski [4], it has found application in numerous areas of machine learning, computer vision, clustering and spatial statistics.

The obvious generalization of the distribution (1) to edge-dependent parameters θ_{ij}, $(i, j) \in E$ is of central interest in such applications, and will be introduced in Section 2.2.2. Let us stress that we follow the statistical mechanics convention of calling (1) an Ising model even if the graph G is not a grid.

In this paper we study the following structural learning problem:

> *Given n i.i.d. samples $\underline{x}^{(1)}$, $\underline{x}^{(2)}$,..., $\underline{x}^{(n)} \in \{+1, -1\}^p$ with distribution $\mu_{G,\theta}(\,\cdot\,)$, reconstruct the graph G.*

For the sake of simplicity, we assume in most of the paper that the parameter θ is known, and that G has no double edges (it is a 'simple' graph). We focus therefore on the key challenge of learning the graph structure associated to the measure $\mu_{G,\theta}(\,\cdot\,)$. This structure is particularly important

* Department of Electrical Engineering, Stanford University

† Departments of Electrical Engineering and Statistics, Stanford University

for extracting the qualitative features of the model, since it encodes its conditional independence properties.

It follows from the general theory of exponential families that, for any $\theta \in (0, \infty)$, the model (1) is identifiable [5]. In particular, the structural learning problem is solvable with unbounded sample complexity and computational resources. The question we address is: for which classes of graphs and values of the parameter θ is the problem solvable under realistic 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

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

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. The general problem is therefore to characterize the functions $n_{\mathsf{Alg}}(G, \theta)$ and $\chi_{\mathsf{Alg}}(G, \theta)$, and to design algorithms that minimize the complexity.

Let us emphasize that these are not the only possible definitions of sample and computational complexity. Alternative definitions are obtained by requiring that the reconstructed structure $\mathsf{Alg}(\underline{x}^{(1)}, \dots, \underline{x}^{(n)})$ is only partially correct. However, for the algorithms considered in this paper, such definitions should not result in qualitatively different behavior[1]

General upper and lower bounds on the sample complexity $n_{\mathsf{Alg}}(G, \theta)$ were proved by Santhanam and Wainwright [6, 7], without however taking into account computational complexity. On the other end of the spectrum, several low complexity algorithms have been developed in the last few years (see Section 1.3 for a brief overview). However the resulting sample complexity bounds only hold under specific assumptions on the underlying model (i.e. on the pair (G, θ)). A general understanding of the trade-offs between sample complexity and computational complexity is largely lacking.

This paper is devoted to the study of the tradeoff between sample complexity and computational complexity for some specific structural learning algorithms, when applied to the Ising model. An important challenge consists in the fact that the model (1) induces subtle correlations between the binary variables $(x_1, \dots, x_p)$. The objective of a structural learning algorithm is to disentangle pairs x_i, x_j that are conditionally independent given the other variables (and hence are not connected by an edge) from those that are instead conditionally dependent (and hence connected by an edge in G). This becomes particularly difficult when θ becomes large and hence pairs x_i, x_j that are not connected by an edge in G become strongly dependent. The next section sets the stage for our work by discussing a simple and concrete illustration of this phenomenon.

1.1 A toy example

As a toy illustration[2] of the challenges of structural learning, we will study the two families of graphs in Figure 1. The two families will be denoted by $\{G_p\}_{p\geq 3}$ and $\{G'_p\}_{p\geq 3}$ and are indexed by the number of vertices p.

Graph G_p has p vertices and $2(p-2)$ edges. Two of the vertices (vertex 1 and vertex 2) have degree $(p-2)$, and $(p-2)$ have degree 2. Graph G'_p has also p vertices, but only one edge between vertices 1 and 2. In other words, graph G'_p corresponds to variables x_1 and x_2 interacting 'directly' (and hence not conditionally independent), while graph G_p describes a situation in which the two variables

[1] Indeed the algorithms considered in this paper reconstruct G by separately estimating the neighborhood of each node i. This implies that any significant probability of error results in a substantially different graph.

[2] A similar example was considered in [8].

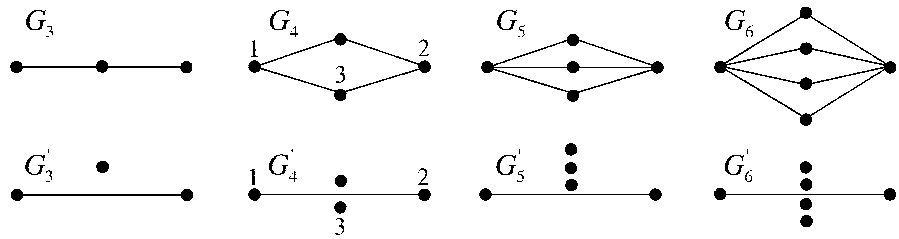


Figure 1: Two families of graphs G_p and G_p' whose distributions $\mu_{G_p,\theta}$ and $\mu_{G_p',\theta'}$ merge as p gets large.

interact 'indirectly' through numerous weak intermediaries (but still are conditionally independent since they are not connected). Fix p, and assume that one of G_p or G_p' is chosen randomly and i.i.d. samples $\underline{x}^{(1)},\dots,\underline{x}^{(n)}$ from the corresponding Ising distribution are given to us.

Can we efficiently distinguish the two graphs, i.e. infer whther the samples were generated using G_p or G_p'? As mentioned above, since the model is identifiable, this task can be achieved with unbounded sample and computational complexity. Further, since model (1) is an exponential family, the $p \times p$ matrix of empirical covariances $(1/n)\sum_{\ell=1}^{n} \underline{x}^{(\ell)}(\underline{x}^{(\ell)})^T$ provides a sufficient statistic for inferring the graph structure.

In this specific example, we assume that different edge strengths are used in the two graphs: θ for graph G_p and θ' for graph G_p' (i.e. we have to distinguish between $\mu_{G_p,\theta}$ and $\mu_{G_p',\theta'}$). We claim that, by properly choosing the parameters θ and θ', we can ensure that the covariances approximately match $|\mathbb{E}_{G_p,\theta}\{x_i x_j\} - \mathbb{E}_{G_p',\theta'}\{x_i x_j\}| = O(1/\sqrt{p})$. Indeed the same remains true for all marginals involving a bounded number of variables. Namely, for all subsets of vertices $U \subseteq [p]$ of bounded size $|\mu_{G_p,\theta}(\underline{x}_U) - \mu_{G_p',\theta'}(\underline{x}_U)| = O(1/\sqrt{p})$. Low-complexity algorithms typically estimate each edge using only a small subset low–dimensional marginal. Hence, they are bound to fail unless the number of samples n diverges with the graph size p. On the other hand, a naive information-theoretic lower bound (in the spirit of [6, 7]) only yields $n_{\mathsf{Alg}}(G,\theta) = \Omega(1)$. This sample complexity is achievable by using global statistics to distinguish the two graphs.

In other words, even for this simple example, a dichotomy emerges: either a number of samples has to grow with the number of parameters, or algorithms have to exploit a large number of marginals of $\mu_{G,\theta}$.

To confirm our claim, we need to compute the covariance of the Ising measures distributions $\mu_{G_p,\theta}$, $\mu_{G_p',\theta'}$. We easily obtain, for the latter graph

$$\mathbb{E}_{G_p',\theta'}\{x_1 x_2\} = \tanh\theta' , \tag{3}$$

$$\mathbb{E}_{G_p',\theta'}\{x_i x_j\} = 0 . \qquad (i,j) \neq (1,2) . \tag{4}$$

The calculation is somewhat more intricate for graph G_p, so we defer complete formulae to Appendix A and report here only the result for $p \gg 1$, $\theta \ll 1$:

$$\mathbb{E}_{G_p,\theta}\{x_1 x_2\} = \tanh\left\{p\theta^2 + O(p\theta^4,\theta)\right\} , \tag{5}$$

$$\mathbb{E}_{G_p,\theta}\{x_i x_j\} = O(\theta, p\theta^3) , \qquad i \in \{1,2\}, j \in \{3,\dots,p\} , \tag{6}$$

$$\mathbb{E}_{G_p,\theta}\{x_i x_j\} = O(\theta^2, p\theta^4) , \qquad i,j \in \{3,\dots,p\} . \tag{7}$$

In other words, variables x_1 and x_2 are strongly correlated (although not connected), while all the

other variables are weakly correlated. By letting $\theta = \sqrt{\theta'/p}$ this covariance structure matches Eqs. (3), (4) up to corrections of order $1/\sqrt{p}$.

Notice that the ambiguity between the two models G_p and G'_p arises because several weak, indirect paths between x_1 and x_2 in graph G_p, add up to the same effect as a strong direct connection. This toy example is hence suggestive of the general phenomenon that strong long-range correlations can 'fake' a direct connection. However, the example is not completely convincing for several reasons:

1. Most algorithms of interest estimate each edge on the basis of a large number of low-dimensional marginals (for instance *all* pairwise correlations).

2. Reconstruction guarantees have been proved for graphs with bounded degree [9, 10, 6, 7, 11], while here we are letting the maximum degree be as large as the system size. Notice however that a the graph considered here are only sparse 'on average'.

3. It may appear that the difficulty in distinguishing graph G_p from G'_p is related to the fact that in the former we take $\theta = O(1/\sqrt{p})$. This is however the natural scaling when the degree of a vertex is large, in order to obtain a non-trivial distribution. If the graph G_p had θ bounded away from 0, this would result in a distribution $\mu_{G_p,\theta}(\underline{x})$ concentrated on the two antipodal configurations: all-$(+1)$ and all-(-1). Structural learning would be equally difficult in this case.

Despite these points, this model provides already a useful counter-example. In Appendix D.3 we will show why, even for bounded p (and hence θ bounded away from 0) the model G_p in Figure 1 'fools' regularized logistic regression algorithm of Ravikumar, Wainwright and Lafferty [11]. Regularized logistic regression reconstructs G'_p instead of G_p.

1.2 Outline of the paper

The rest of this paper is devoted to bounding the sample complexity n_{Alg} and computational complexity χ_{Alg} for a number of graph models, as a function of θ. Results of this analysis are presented in Section 2 for three algorithms: a simple thresholding algorithm, the conditional independence test method of [10] and the penalized pseudo-likelihood method of [11]. In Section 3, we validate our analysis through numerical simulations. Finally, Section 4 contains the proofs with some technical details deferred to the appendices.

This analysis unveils a general pattern: *when the model (1) develops strong correlations, several low-complexity algorithms fail, or require a large number of samples.* What does 'strong correlations' mean? As the toy example in the previous section demonstrates, correlations arise from a trade-off between the degree (which we will characterize here via the maximum degree Δ), and the interaction strength θ. It can be ascribed to a few strong connections (large θ) or to a large number of weak connections (large Δ). Is there any meaningful way to compare and combine these quantities (θ and Δ)? An answer is suggested by the theory of Gibbs measures which predicts a dramatic change of behavior when θ crosses the so-called 'uniqueness threshold' $\theta_{\rm uniq}(\Delta) = \operatorname{atanh}(1/(\Delta-1))$ [12]. For $\theta < \theta_{\rm uniq}(\Delta)$ Gibbs sampling mixes rapidly and far apart variables in G are roughly independent [13]. Vice versa, for any $\theta > \theta_{\rm uniq}(\Delta)$ there exist graph families on which Gibbs sampling is slow, and far apart variables are strongly dependent [14]. While polynomial sampling algorithms exists for all $\theta > 0$ [15], for $\theta < 0$, in the regime $|\theta| > \theta_{\rm uniq}(\Delta)$ sampling is arguably #-P hard [16]. Related to the uniqueness threshold is also the phase transition threshold, which is graph dependent, with typically $\theta_{\rm crit} \le \text{const.}/\Delta$.

We will see that this is indeed a relevant way of comparing interaction strength and degree, even for structural learning. Al the algorithms we analyzed (mentioned above) provably fail for

$\theta \gg \text{const.}/\Delta$, for a number of 'natural' graph families. Our work raises several fascinating questions, the most important being the construction of structural learning algorithm with provable performance guarantees in the strongly dependent regime $\theta_{\rm crit} \gg \text{const.}/\Delta$. The question as to whether such an algorithm exists is left open by the present paper (but see next section for an overview of earlier work).

Let us finally emphasize that we do not think that any of the specific families of graphs studied in the present paper is intrinsically 'hard' to learn. For instance, we show below that the regularized logistic regression method of [11] fails on random regular graphs, while it is easy to learn such graphs using the simple thresholding algorithm of Section 2.1. The specific families where indeed chosen mostly because they are analytically tractable.

1.3 Further related work

Traditional algorithms for learning Ising models were developed in the context of Boltzmann machines [4, 17, 18]. These algorithms try to solve the maximum likelihood problem by gradient ascent. Estimating the gradient of the log-likelihood function requires to compute expectations with respect to the Ising distribution. In these works, this was done using the Markov Chain Monte Carlo (MCMC) method, and more specifically Gibbs sampling.

We shall not consider this approach in our study for two type of reasons. First of all, it does not output a 'structure' (i.e. a sparse subset of the $\binom{p}{2}$ potential edges): because of approximation errors, it yields non-zero values for all the edges. This problem can in principle be overcome by using suitably regularized objective functions, but such a modified algorithm was never studied.

Second, the need to compute expectation values with respect to the Ising distribution, and the use of MCMC to achieve this goal, poses some fundamental limitations. As mentioned above, the Markov chain commonly used by these methods is simple Gibbs sampling. This is known to have mixing time that grows exponentially in the number of variables for $\theta > \theta_{\rm uniq}(\Delta)$, and hence does not yield good estimates of the expectation values in practice. While polynomial sampling schemes exist for models with $\theta > 0$ [15], they do not apply to $\theta < 0$ or to general models with edge-dependent parameters θ_{ij}. Already in the case $\theta < 0$, estimating expectation values of the Ising distribution is likely to be #-P hard [16].

Abbeel, Koller and Ng [9] first developed a method with computational complexity provably polynomial in the number of variables, for bounded maximum degree, and logarithmic sample complexity. Their approach is based on ingenious use of the Hammersley-Clifford representation of Markov Random Fields. Unfortunately, the computational complexity of this approach is of order $p^{\Delta+2}$ which becomes unpractical for reasonable values of the degree and network size (and superpolynomial for Δ diverging with p). The algorithm by Bresler, Mossel and Sly [10] studied in Section 2.2.1 presents similar limitations, that the authors overcome (in the small θ regime) by exploiting the correlation decay phenomenon.

An alternative point of view consists in using standard regression methods. In the context of Ising models, Ravikumar, Wainwright and Lafferty [11] showed that the neighborhood of a vertex i can be efficiently reconstructed by solving an appropriate regularized regression problem. More precisely, the values of variable x_i are regressed against the value of all the other variables. The logistic regression log-likelihood is regularized by adding an ℓ_1-penalty that promotes the selection of sparse graph structures. We will analyze this method in Section 2.2.2. The approach of [11] extends to non-Gaussian models earlier work by Meinshausen and Bühlmann [19]. Let us notice in passing that the case of Gaussian graphical models is substantially easier since the log-likelihood of a given model can be evaluated easily in this case [20].

A short version of this paper was presented at the 2009 Neural Information Processing Systems

symposium. Since then, at least two groups explored the challenges put forward in our work. Anandkumar, Tan and Willsky [21] prove that, for sequences of random graphs which are sparse on average (i.e. with bounded average degree), structural learning is possible throughout the correlation decay regime $\theta < \theta_{\rm crit}$. This result generalizes our analysis of random regular graphs (see next section), to the more challenging case of graphs with random degrees. Cocco and Monasson [22] proposed and 'adaptive cluster' heuristics and demonstrated empirically good performances for specific graph families, also for $\theta > \theta_{\rm crit}$. A mathematical analysis of their approach is lacking.

2 Results

2.1 The simple thresholding algorithm

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

$$\widehat{C}_{ij} \equiv \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)}\,, \tag{8}$$

for $i, j \in V$.

THRESHOLDING(samples $\{x^{(\ell)}\}$, threshold τ)
Compute the empirical correlations $\{\widehat{C}_{ij}\}_{(i,j)\in V\times V}$;
For each $(i,j) \in V \times V$
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 (for proofs see Section 4.2).

Theorem 2.1. *If G is a tree, and $\tau(\theta) = (\tanh\theta + \tanh^2\theta)/2$, then*

$$n_{\mathsf{Thr}(\tau)}(G,\theta) \leq \frac{32}{(\tanh\theta - \tanh^2\theta)^2}\,\log\frac{2p}{\delta}\,. \tag{9}$$

Theorem 2.2. *If G has maximum degree $\Delta > 1$ and if $\theta < \operatorname{atanh}(1/(2\Delta))$ then there exists $\tau = \tau(\theta)$ such that*

$$n_{\mathsf{Thr}(\tau)}(G,\theta) \leq \frac{32}{(\tanh\theta - \frac{1}{2\Delta})^2}\,\log\frac{2p}{\delta}\,. \tag{10}$$

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

Theorem 2.3. *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. Hence trees can be learnt from $O(\log p)$ samples irrespectively of their topology and maximum degree (assuming $\theta \neq \infty$). The same happens on bounded-degree graphs if $\theta \leq$ const.$/\Delta$. However, for $\theta >$ const.$/\Delta$, there exists families of bounded degree graphs with long-range correlations.

2.2 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 the strength of correlations. We focus on two type of algorithms and only include the proof of our most challenging result, Theorem 2.8 (for the proof see Section 4.3).

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: $|\partial i| \le \Delta$.

2.2.1 Local Independence Test

A recurring approach to structural learning consists in exploiting the conditional independence structure encoded by the graph [9, 10, 23, 24].

Let us consider, to be definite, the approach of [10], 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[3]: $\mu_{G,\theta}(x_r|\underline{x}_{\partial r})$. 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, $|W| \le \Delta$. On the other hand, if $U \nsubseteq \partial r$, then it is possible to find W (with $|W| \le \Delta$) 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 $\textsc{Score}(U) > \epsilon/2$;
3:	Repeat for all nodes $r \in V$;

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

$$\min_{W,j}\ \max_{x_i,\underline{x}_W,\underline{x}_U,x_j} |\widehat{\mathbb{P}}_{n,G,\theta}\{X_i = x_i|\underline{X}_W = \underline{x}_W, \underline{X}_U = \underline{x}_U\} - \widehat{\mathbb{P}}_{n,G,\theta}\{X_i = x_i|\underline{X}_W = \underline{x}_W, \underline{X}_{U\backslash j} = \underline{x}_{U\backslash j}, X_j = x_j\}|\,. \tag{11}$$

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

$$\begin{aligned}&\widehat{\mathbb{P}}_{n,G,\theta}\{\underline{X}_W = \underline{x}_W, \underline{X}_U = \underline{x}_U\} > \gamma/2\\ &\widehat{\mathbb{P}}_{n,G,\theta}\{\underline{X}_W = \underline{x}_W, \underline{X}_{U\backslash j} = \underline{x}_{U\backslash j}, X_j = x_j\} > \gamma/2\end{aligned} \tag{12}$$

$\widehat{\mathbb{P}}_{n,G,\theta}$ is the empirical distribution calculated from the samples $\{\underline{x}^{(\ell)}\}_{\ell=1}^n$. 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 $\textsc{Score}(U)$ 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 [10], hence omitted.

[3] 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.

Theorem 2.4. *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*

$$n_{\mathsf{Ind}(\epsilon,\gamma)}(G,\theta) \leq \frac{100\Delta}{\epsilon^2\gamma^4} \log \frac{2p}{\delta}\,, \tag{13}$$

$$\chi_{\mathsf{Ind}(\epsilon,\gamma)}(G,\theta) \leq K\,(2p)^{2\Delta+1} \log p\,. \tag{14}$$

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 [10]. The idea is to identify a set of 'potential neighbors' of vertex r via thresholding:

$$B(r) = \{i \in V : \widehat{C}_{ri} > \kappa/2\}\,. \tag{15}$$

For each node $r \in V$, we evaluate $\textsc{Score}(U)$ by restricting the minimum in Eq. (11) 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 low temperature.

Theorem 2.5. *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*

$$n_{\mathsf{IndD}(\epsilon,\gamma,\kappa)}(G,\theta) \leq 8(\kappa^2 + 8^\Delta) \log \frac{4p}{\delta}\,, \tag{16}$$

$$\chi_{\mathsf{IndD}_{(\epsilon,\gamma,\kappa)}}(G,\theta) \leq K' p \Delta^{\Delta \frac{\log(4/\kappa)}{\alpha}} + K'\Delta p^2 \log p\,. \tag{17}$$

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

2.2.2 Regularized Pseudo-Likelihoods

A different approach to the learning problem consists in maximizing an appropriate empirical likelihood function [11, 25, 26, 27, 19, 28]. In order to control statistical fluctuations, and select sparse graphs, a regularization term is often added to the cost function.

As a specific low complexity implementation of this idea, we consider the ℓ_1-regularized pseudo-likelihood method of [11]. For each node r, the following likelihood function is considered

$$L(\underline{\theta}; \{x^{(\ell)}\}) = -\frac{1}{n}\sum_{\ell=1}^{n} \log \mathbb{P}_{n,G,\underline{\theta}}(x_r^{(\ell)} | x_{\setminus r}^{(\ell)}) \tag{18}$$

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

$$\mu_{G,\underline{\theta}}(\underline{x}) = \frac{1}{Z_{G,\underline{\theta}}} \prod_{i,j\in V} e^{\theta_{ij} x_i x_j} \tag{19}$$

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 2.6. *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*

$$n_{\mathsf{Rlr}(\lambda)}(G,\theta) \leq K_2\, \theta^{-2}\, \Delta\, \log \frac{8p^2}{\delta}\,. \tag{20}$$

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 [11] (denoted there by A1 and A2) hold, and then computing the probability of success as a function of n with slightly more care. The details of the proof are written in Appendix B.

In order to prove a converse to the above result, we need to make some assumptions on λ.

Definition 2.7. *Given $\theta > 0$, we say that λ is* reasonable *for that value of θ if the following conditions hold: (i) $\mathsf{Rlr}(\lambda)$ is successful with probability larger than $1/2$ on any star graphs (a graph composed by a vertex r connected to Δ neighbors, plus isolated vertices) if n is chosen sufficiently high; (ii) $\lambda \leq \delta(n)$ for some sequence $\delta(n) \downarrow 0$.*

In other words, assumption (i) requires the algorithm to be successful on a particularly simple class of graphs, and hence does not entail any loss of generality. Assumption (ii) encodes instead the standard way of scaling regularization terms, by letting them vanish as the number of samples increases. This is necessary in order to get asymptotic consistency of the parameter values θ_{ij}. With these assumptions we can state the following converse theorem, whose proof is deferred to Section 4.3.

Theorem 2.8. *There exists a numerical constant K such that the following happens. If $\theta > K/\Delta$,$\Delta > 3$, 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, as part of the proof of Theorem 2.8, and as proved in Appendix D, we have the following facts about $\mathsf{Rlr}(\lambda)$:

- If G is a tree, then $\mathsf{Rlr}(\lambda)$ recover G with high probability for any θ (for a suitable λ);
- For every graph G_p in the family described in Section 1.1, $\mathsf{Rlr}(\lambda)$ fails with high probability for θ large enough and for all λ;
- If G is sampled uniformly from the ensemble of regular graphs $\mathsf{Rlr}(\lambda)$ fails with high probability for θ large enough and λ 'reasonable';
- if G is a large two dimensional grid It fails with high probability for θ large enough and λ 'reasonable'.

We note here that Theorem 2.8 relies on proving that a so-called 'Incoherence condition' is necessary for Rlr to successfully reconstruct G. Although a similar result was proven in [29] for model selection using the Lasso, this paper is the first to prove that a similar Incoherence condition is also necessary when the underlying model is the Ising model.

The intuition behind this is quite simple. Begin by noticing that when $n \to \infty$, and under the restriction that $\lambda \to 0$, solutions given by Rlr converge to $\underline{\theta}^*$ as $n \to \infty$ [11]. Hence, for large n, we can expand L in a quadratic function centered around $\underline{\theta}^*$ plus a small stochastic error term. Consequently, when adding the regularization term to L, we obtain cost function analogous to the Lasso plus an error term that needs to be controlled. The study of the dominating contribution leads to the incoherence condition.

In general there are no practical ways to evaluate the incoherence condition for a given graphical model. This requires in fact to compute expectations with respect to the Ising distribution. As discussed above, this is hard for $|\theta| > \theta_{\rm uniq}(\Delta)$. Hence this condition was not checked for families of graphs. A large part of our technical contribution consists indeed in filling this gap. To this end, we use tools from mathematical statistical mechanics, namely low temperature series for Ising models on grids [30, 31], and local weak convergence results for Ising models on random graphs [32, 33].

3 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 $\mathsf{Rlr}(\lambda)$. Among other learning algorithms, $\mathsf{Rlr}(\lambda)$ 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 $\theta \geq \theta_{\rm uniq}$, 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. Notice that $\mathsf{Rlr}(\lambda)$ had been tested in [11] only on tree graphs G, or in the weakly coupled regime $\theta < \theta_{\rm uniq}$. In these cases sampling from the Ising model is easy, but structural learning is also intrinsically easier.

Figure reports the success probability of $\mathsf{Rlr}(\lambda)$ 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 $\mathsf{Rlr}(\lambda)$ to each vertex of 8 graphs (thus averaging over 392 runs of $\mathsf{Rlr}(\lambda)$), 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 [11] and is empirically the most satisfactory), and searched over λ_0.

The data clearly illustrate the phenomenon discussed in the previous pages. 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 $\theta_{\rm crit}(\rho = 0.3) \approx 0.7$ [34, 35].

Figure 3 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 4.3 below, and get $\theta_{\rm thr}(\Delta = 4) \approx 0.4203$, which compares favorably with the data.

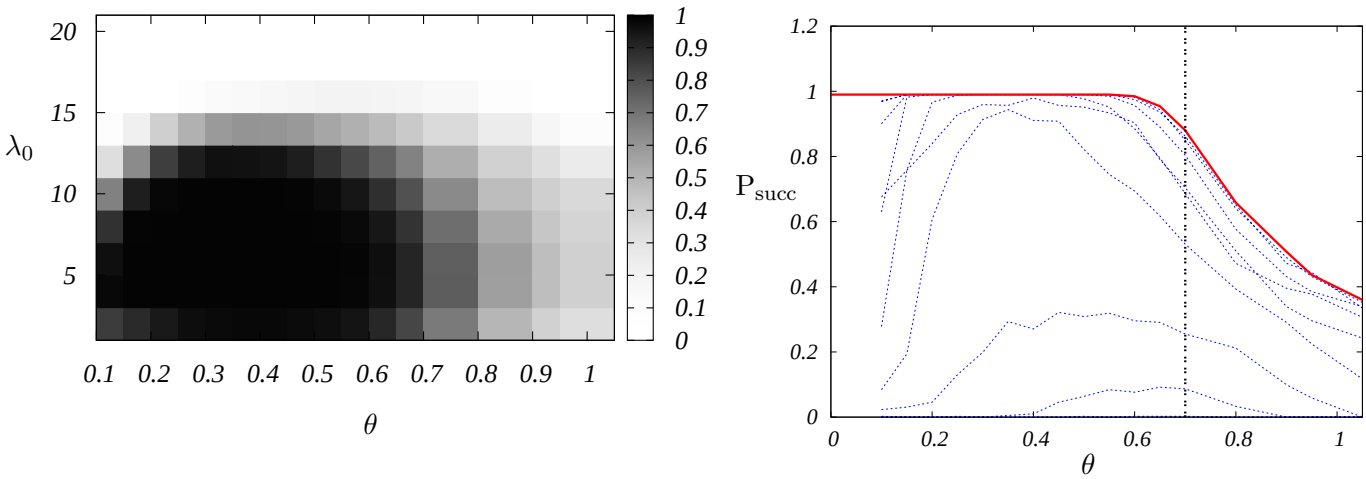


Figure 2: Learning random subgraphs of a 7×7 ($p = 49$) two-dimensional grid from $n = 4500$ Ising models samples, using regularized logistic regression. Left: success probability as a function of the model parameter θ and of the regularization parameter λ_0 (darker corresponds to highest probability). 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.

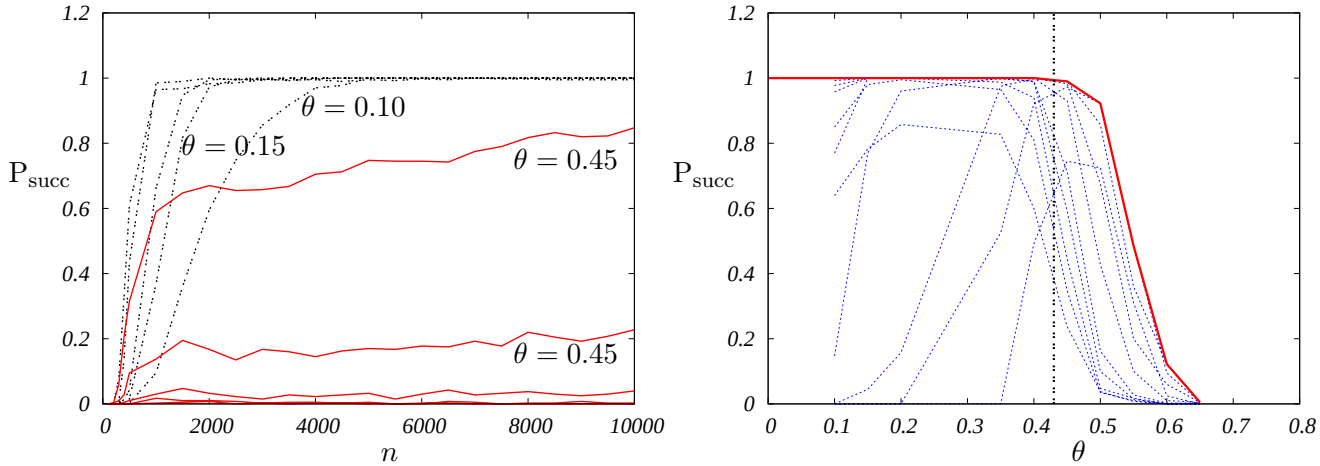


Figure 3: Learning uniformly random graphs of degree $\Delta = 4$ from Ising models samples, using regularized logistic regression. Left: success probability as a function of the number of samples n for several values of θ. Dotted: $\theta = 0.10$, 0.15, 0.20, 0.35, 0.40 (in all these cases $\theta < \theta_{\rm thr}(\Delta = 4)$). Dashed: $\theta = 0.45$, 0.50, 0.55, 0.60, 0.65 ($\theta > \theta_{\rm thr}(4)$, some of these are indistinguishable from the axis). Right: the same data plotted for several choices of λ versus θ as in Fig. 2, right panel.

4 Proofs

4.1 Notation and important remarks

Before proceeding it is convenient to introduce some notation and make some important remarks. If V is a matrix and R is an index set then V_R denotes the vector formed by all entries whose index lies in R and similarly, 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 before, we let r be the vertex whose neighborhood we are trying to reconstruct and define $S = \partial r$ and $S^c = V \setminus \partial r \cup r$. Since the cost function $L(\underline{\theta}; \{\underline{x}^{(\ell)}\}) + \lambda \|\underline{\theta}\|_1$ only depends 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 $\mathsf{Rlr}(\lambda)$ when the number of samples is n. Note that, since we assumed $\underline{\theta}^* \geq 0$, we have $\underline{\theta}^*_S > 0$ and hence $\hat{z}^*_S = \mathbb{1}$. 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 $Q(\underline{\theta})$ exists a.s. and is the Hessian of $\mathbb{E}_{G,\underline{\theta}} \log \mathbb{P}_{G,\underline{\theta}}(X_r | X_{\setminus r})$ where $\mathbb{E}_{G,\underline{\theta}}$ is the expectation with respect to (19) and $\underline{X}$ is a random variable distributed according to (19). It is convenient to recall here the expressions for the Hessian and gradient of L for finite n and in the limit when $n \to \infty$. For all $i, j \in V \setminus \{r\}$ we have,

$$Q^n_{ij}(\underline{\theta}) = \frac{1}{n} \sum_{\ell=1}^{n} \frac{x_i^{(\ell)} x_j^{(\ell)}}{\cosh^2\left(\sum_{t\in V\setminus\{r\}} \theta_{rt} x_t^{(\ell)}\right)}, \tag{21}$$

$$Q_{ij}(\underline{\theta}) = \mathbb{E}_{G,\underline{\theta}^*}\left(\frac{X_i X_j}{\cosh^2(\sum_{t\in V\setminus\{r\}} \theta_{rt} X_t)}\right), \tag{22}$$

$$[\nabla L^n(\underline{\theta})]_i = \frac{1}{n} \sum_{\ell=1}^{n} x_i^{(\ell)} \Big(\tanh\Big(\sum_{t\in V\setminus\{r\}} \theta_{rt} x_t^{(\ell)} \Big) - x_r^{(\ell)} \Big), \tag{23}$$

$$[\nabla L(\underline{\theta})]_i = \mathbb{E}_{G,\underline{\theta}^*}\Big\{ X_i \tanh\Big(\sum_{t\in V\setminus\{r\}} \theta_{rt} X_t \Big) \Big\} - \mathbb{E}_{G,\underline{\theta}^*}\{X_i X_r\}. \tag{24}$$

Note that from the last expression it follows that $\nabla L(\underline{\theta}^*) = 0$.

We will denote the maximum and minimum eigenvalue of a symmetric matrix M by $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ respectively. Recall that $\|M\|_\infty = \max_i \sum_j |M_{ij}|$.

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. When clear from the context and since all the examples that we work on have $\theta_{ij} \in \{0, \theta\ \}$, we will write $\mathbb{E}_{G,\theta}$ as $\mathbb{E}_{G,\underline{\theta}}$ or even simply $\mathbb{E}$. Similarly, $\mathbb{P}_{G,\underline{\theta}}$ will be sometimes written as simply $\mathbb{P}_{G,\theta}$ or just $\mathbb{P}$. A subscript n under $\mathbb{P}_{G,\underline{\theta}}$, i.e. $\mathbb{P}_{n,G,\underline{\theta}}$, will be introduced to denote the product measure formed by n copies of model (19). Through out this section $\mathrm{P}_{\mathrm{succ}}$ will denote the probability of success of a given algorithm, that is, the probability that the algorithm is able to recover the underlying G exactly.

Throughout this section G is a graph of maximum degree Δ.

4.2 Simple Thresholding

In the following we let $C_{ij} \equiv \mathbb{E}_{G,\theta}\{X_i X_j\}$ where expectation is taken with respect to the Ising model (1).

Proof. (Theorem 2.1) If G is a tree then $C_{ij} = \tanh\theta$ for all $(ij) \in E$ and $C_{ij} \le \tanh^2\theta$ for all $(ij) \notin E$. To see this notice that only paths that connect i to j contribute to C_{ij} and given that G is a tree there is only one such path and its length is exactly 1 if $(i,j) \in E$ and at least 2 when $(i,j) \notin E$. The probability that $\mathsf{Thr}(\tau)$ fails is

$$1 - \mathrm{P}_{\rm succ} = \mathbb{P}_{n,G,\theta}\{\widehat{C}_{ij} < \tau \text{ for some } (i,j) \in E \text{ or } \widehat{C}_{ij} \ge \tau \text{ for some } (i,j) \notin E\}\,. \tag{25}$$

Let $\tau = (\tanh\theta + \tanh^2\theta)/2$. Applying Azuma-Hoeffding inequality to $\widehat{C}_{ij} = \frac{1}{n}\sum_{\ell=1}^n x_i^{(\ell)} x_j^{(\ell)}$ we have that if $(i,j) \in E$ then,

$$\mathbb{P}_{n,G,\theta}(\widehat{C}_{ij} < \tau) = \mathbb{P}_{n,G,\theta}\left(\sum_{\ell=1}^n (x_i^{(\ell)} x_j^{(\ell)} - C_{ij}) < n(\tau - \tanh\theta)\right) \le e^{-\frac{1}{32}n(\tanh\theta - \tanh^2\theta)^2} \tag{26}$$

and if $(i,j) \notin E$ then similarly,

$$\mathbb{P}_{n,G,\theta}(\widehat{C}_{ij} \ge \tau) = \mathbb{P}_{n,G,\theta}\left(\sum_{\ell=1}^n (x_i^{(\ell)} x_j^{(\ell)} - C_{ij}) \ge n(\tau - \tanh^2\theta)\right) \le e^{-\frac{1}{32}n(\tanh\theta - \tanh^2\theta)^2}. \tag{27}$$

Applying union bound over the two possibilities, $(i,j) \in E$ or $(i,j) \notin E$, and over the edges ($|E| < p^2/2$), we can bound $\mathrm{P}_{\rm succ}$ by

$$\mathrm{P}_{\rm succ} \ge 1 - p^2\, e^{-\frac{1}{32}n(\tanh\theta - \tanh^2\theta)^2}\,. \tag{28}$$

Imposing the right hand side to be larger than δ proves our result. □

Proof. (Theorem 2.2) We will prove that, for $\theta < \operatorname{arctanh}(1/(2\Delta))$, $C_{ij} \ge \tanh\theta$ for all $(i,j) \in E$ and $C_{ij} \le 1/(2\Delta)$ for all $(ij) \notin E$. In particular $C_{ij} < C_{kl}$ for all $(i,j) \notin E$ and all $(k,l) \in E$. The theorem follows from this fact via union bound and Azuma-Hoeffding inequality as in the proof of Theorem 2.1.

The bound $C_{ij} \ge \tanh\theta$ for $(ij) \in E$ is a direct consequence of Griffiths inequality [36] : compare the expectation of $x_i x_j$ in G with the same expectation in the graph that only includes edge (i,j).

The second bound is derived using the technique of [35], i.e., bound C_{ij} by the generating function for self-avoiding walks on the graphs from i to j. More precisely, assume $l = \mathrm{dist}(i,j)$ and denote by $N_{ij}(k)$ the number of self avoiding walks of length k between i and j on G. Then [35] proves that

$$C_{ij} \le \sum_{k=l}^{\infty} (\tanh\theta)^k N_{ij}(k) \le \sum_{n=l}^{\infty} \Delta^{k-1}(\tanh\theta)^k \le \frac{\Delta^{l-1}(\tanh\theta)^l}{1 - \Delta\tanh\theta} \le \frac{\Delta(\tanh\theta)^2}{1 - \Delta\tanh\theta}. \tag{29}$$

If $\theta < \operatorname{arctanh}(1/(2\Delta))$ the above implies $C_{ij} \le 1/(2\Delta)$ which is our claim. □

Proof. (Theorem 2.3) The theorem is proved by constructing G as follows: sample a uniformly random regular graph of degree Δ over the $p-2$ vertices $\{1, 2, \dots, p-2\} \equiv [p-2]$. Add an extra edge between nodes $p-1$ and p. The resulting graph is not connected. We claim that for $\theta > K/\Delta$ and with probability converging to 1 as $p \to \infty$, there exist $i, j \in [p-2]$ such that $(i,j) \notin E$ and $C_{ij} > C_{p-1,p}$. As a consequence, thresholding fails.

Obviously $C_{p-1,p} = \tanh\theta$. Choose $i \in [p-2]$ uniformly at random, and j a node at a fixed distance t from i. We can compute C_{ij} as $p \to \infty$ using the same local weak convergence result as in the proof of Lemma 4.3. Namely, C_{ij} converges to the correlation between the root and a leaf node in the tree Ising model (45). In particular one can show, [33], that

$$\lim_{p\to\infty} C_{ij} \ge m(\theta)^2\,, \tag{30}$$

where $m(\theta) = \tanh(\Delta h^*/(\Delta - 1))$ and h^* is the unique positive solution of Eq. (46).

The proof is completed by showing that $\tanh\theta < m(\theta)^2$ for all $\theta > K/\Delta$. □

4.3 Proof of Theorem 2.8: failure of regularized logistic regression

In order to prove Theorem 2.8, we need a few auxiliary results. Our first auxiliary results establishes that, if λ is small, then $\|Q^*_{S^cS}{Q^*_{SS}}^{-1}\hat{z}^*_S\|_\infty > 1$ is a sufficient condition for the failure of $\mathsf{Rlr}(\lambda)$. We recall here that the subgradient of $\|\underline{\theta}\|_1$ evaluated at $\underline{\theta}^*$,that is $\hat{z}^*$, satisfies $\hat{z}^*_S = \mathbb{1}$.

Lemma 4.1. *Assume $[Q^*_{S^cS}{Q^*_{SS}}^{-1}\hat{z}^*_S]_i \geq 1+\epsilon$ for some $\epsilon > 0$ and some row $i \in V$, $\sigma_{\min}(Q^*_{SS}) \geq C_{\min} > 0$, and $\lambda < C^3_{\min}\epsilon/(2^7(1+\epsilon^2)\Delta^3)$. Then the success probability of $\mathsf{Rlr}(\lambda)$ is upper bounded as*

$$\mathrm{P}_{\mathrm{succ}} \leq 4\Delta^2 e^{-n\delta_A^2} + 4\Delta\, e^{-n\lambda^2\delta_B^2} \tag{31}$$

where $\delta_A = (C^2_{\min}/32\Delta)\epsilon$ and $\delta_B = (C_{\min}/64\sqrt{\Delta})\epsilon$.

The next Lemma implies that, for λ to be 'reasonable' (in the sense introduced in Section 2.2.2), $n\lambda^2$ must be unbounded with respect to p. In fact, by this lemma, if we choose n to be very large and choose a sequence of star graphs of increasing number of nodes but with only one edge between the central node and the remaining nodes, then, unless K is increasing with p, $\mathsf{Rlr}(\lambda)$ will fail to reconstruct the graph with a probability greater than 1/2, which is a contradiction if λ is 'reasonable'.

Lemma 4.2. *There exist $M = M(K,\theta) > 0$ decreasing with K for $\theta > 0$ such that the following is true: If G is the (star) graph with vertex set $V = [p]$ and edge set $E = \{(r,i)\}$ (e.g. $r = 1$, $i = 2$) and $n\lambda^2 \leq K$, then*

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

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 [32].

Lemma 4.3. *Let G be a uniformly random regular graph of degree $\Delta > 3$. Then, there exists $\theta_{\mathrm{thr}}(\Delta)$ such that, for $\theta > \theta_{\mathrm{thr}}(\Delta)$, $\|Q^*_{S^cS}{Q^*_{SS}}^{-1}\hat{z}^*_S\|_\infty \geq 1+\epsilon(\theta,\Delta)$ with probability converging to 1 as $p \to \infty$ ($\epsilon(\theta,\Delta) > 0$ and $\epsilon(\theta,\Delta) \to 0$ as $\theta \to \infty$).*

Furthermore, for large Δ, $\theta_{\mathrm{thr}}(\Delta) = \tilde{\theta}\,\Delta^{-1}(1+o(1))$. The constant $\tilde{\theta}$ is given by $\tilde{\theta} = h^2_\infty$ and h_∞ is the unique positive solution of

$$h_\infty \tanh h_\infty = 1. \tag{33}$$

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 4.1, 4.2 and 4.3 are sketched in the next subsection.

Proof. (Theorem 2.8) 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 4.3, for any p large enough we can choose a Δ-regular graph $G_p = (V = [p], E_p)$ and vertex $r \in V$ such that $|Q^*_{S^cS}{Q^*_{SS}}^{-1}\hat{z}^*_S|_i > 1+\epsilon$ for some $i \in V \setminus r$ (Indeed most vertices r and graphs G_p will work).

By Theorem 1 in [10] we can assume without loss of generality $n > K'\Delta\log p$ for some small constant K'. Further by Lemma 4.2, $n\lambda^2 \geq F(p)$ for some $F(p) \uparrow \infty$ as $p \to \infty$ and the condition of Lemma 4.1 on λ is satisfied since by the assumption that λ is 'reasonable' we have $\lambda \to 0$ as $n \to \infty$. Using these results in Eq. (31) of Lemma 4.1 we get the following upper bound on the success probability

$$\mathrm{P}_{\mathrm{succ}}(G_p) \leq 4\Delta^2 p^{-\delta_A^2 K'\Delta} + 2\Delta\, e^{-F(p)\delta_B^2}\,. \tag{34}$$

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

4.3.1 Proofs of auxiliary lemmas

Proof. (Lemma 4.1) This proof follows closely the proof of Proposition 1 in [11]. For a matter of clarify of exposition we will include all the steps, even if these do not differ from the exposition done in [11].

We will show that (under the assumptions of the Lemma on the Incoherence Condition, $\sigma_{\min}(Q^*_{SS})$ and λ) if $\hat{\underline{\theta}} = (\hat{\underline{\theta}}_S, \hat{\underline{\theta}}_{S^C}) = (\hat{\underline{\theta}}_S, 0)$ with $\hat{\underline{\theta}}_S > 0$ then the probability that $\mathsf{Rlr}(\lambda)$ returns $\hat{\underline{\theta}}$ is upper bounded as in Eq. (31). More specifically, we will show that this $\hat{\underline{\theta}}$ will not satisfy the stationarity condition $\nabla L(\hat{\underline{\theta}}) + \lambda \hat{z} = 0$ with high probability for any subgradient $\hat{z}$ of the function $\|\underline{\theta}\|_1$ at $\hat{\underline{\theta}}$.

To simplify notation we will omit $\{\underline{x}^{(\ell)}\}$ in all the expressions involving and derived from L.

Assume the event $\nabla L(\hat{\underline{\theta}}) + \lambda \hat{z} = 0$ holds for some $\hat{\underline{\theta}}$ as specified above. An application of the mean value theorem yields

$$\nabla^2 L(\underline{\theta}^*)[\hat{\underline{\theta}} - \underline{\theta}^*] = W^n - \lambda \hat{z} - R^n \,, \tag{35}$$

where $W^n = -\nabla L(\underline{\theta}^*)$ and $[R^n]_j = [\nabla^2 L(\bar{\underline{\theta}}^{(j)}) - \nabla^2 L(\underline{\theta}^*)]^T_j (\hat{\underline{\theta}} - \underline{\theta}^*)$ 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 write

$$Q^n_{S^C S}(\hat{\underline{\theta}}_S - \underline{\theta}^*_S) = W^n_{S^C} - \lambda \hat{z}_{S^C} + R^n_{S^C}, \tag{36}$$

$$Q^n_{SS}(\hat{\underline{\theta}}_S - \underline{\theta}^*_S) = W^n_S - \lambda \hat{z}_S + R^n_S. \tag{37}$$

Eliminating $\hat{\underline{\theta}}_S - \underline{\theta}^*_S$ from the two expressions we obtain

$$[W^n_{S^C} - R^n_{S^C}] - Q^n_{S^C S}(Q^n_{SS})^{-1}[W^n_S - R^n_S] + \lambda Q^n_{S^C S}(Q^n_{SS})^{-1}\hat{z}_S = \lambda \hat{z}_{S^C} \,. \tag{38}$$

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

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

Recalling that $\hat{z}_S = \mathbb{1}$ and using the above decomposition we can lower bound the absolute value of the indexed-i component of $\hat{z}_{S^C}$ by

$$|\hat{z}_i| \geq \|[Q^*_{S^C S}{Q^*_{SS}}^{-1}\hat{z}_S]_i\|_\infty - \|T_{1,i}\|_1 - \|T_{2,i}\|_1 - \|T_{3,i}\|_1 \tag{39}$$
$$-\Big|\frac{W^n_i}{\lambda}\Big| - \Big|\frac{R^n_i}{\lambda}\Big| - \|[Q^n_{S^C S}(Q^n_{SS})^{-1}]_i\| \left(\Big\|\frac{W^n_S}{\lambda}\Big\|_\infty + \Big\|\frac{R^n_S}{\lambda}\Big\|_\infty\right).$$

We will now assume that the samples $\{x^{(\ell)}\}$ are such that the following event holds (notice that $i \in S^C$),

$$\mathcal{E}_i \equiv \left\{ \|Q^n_{S\cup\{i\}\, S} - Q^*_{S\cup\{i\}\, S}\|_\infty < \xi_A, \Big\|\frac{W^n_{S\cup\{i\}}}{\lambda}\Big\|_\infty < \xi_B \right\}, \tag{40}$$

where $\xi_A \equiv C^2_{\min}\epsilon/(8\Delta)$ and $\xi_B \equiv C_{\min}\epsilon/(16\sqrt{\Delta})$.

From relations (21) to (24) in Section 4.1 we know that $\mathbb{E}_{G,\theta}(Q^n) = Q^*$, $\mathbb{E}_{G,\theta}(W^n) = 0$ and that both $Q^n - Q^*$ and W^n are sums i.i.d. random variables bounded by 2. From this, a simple application of Azuma-Hoeffding inequality yields [4].

$$\mathbb{P}_{n,G,\theta}(|Q^n_{ij} - Q^*_{ij}| > \delta) \leq 2e^{-\frac{\delta^2 n}{8}}, \tag{41}$$

$$\mathbb{P}_{n,G,\theta}(|W^n_{ij}| > \delta) \leq 2e^{-\frac{\delta^2 n}{8}}, \tag{42}$$

[4] For full details see the proof of Lemma 2 and the discussion following Lemma 6 in [11]

for all i and j. Applying union bound we conclude that the event $\mathcal{E}_i$ holds with probability at least

$$1 - 2\Delta(\Delta+1)e^{-\frac{n\xi_A^2}{8}} - 2(\Delta+1)e^{-\frac{n\lambda^2\xi_B^2}{8}} \geq 1 - 4\Delta^2 e^{-n\delta_A^2} - 4\Delta e^{-n\lambda^2\delta_B}, \tag{43}$$

where $\delta_A = C_{\min}^2\epsilon/(32\Delta)$ and $\delta_B = C_{\min}\epsilon/(64\sqrt{\Delta})$.

If the event $\mathcal{E}_i$ holds then $\sigma_{\min}(Q^n_{SS}) > \sigma_{\min}(Q^*_{SS}) - C_{\min}/2 > C_{\min}/2$. Since $\|[Q^n_{S^cS}(Q^n_{SS})^{-1}]_i\|_\infty \leq \|{Q^n_{SS}}^{-1}\|_2\|Q^n_{Si}\|_2$ and $|Q^n_{ji}| \leq 1 \forall i,j$ we can write $\|[Q^n_{S^cS}(Q^n_{SS})^{-1}]_i\|_\infty \leq 2\sqrt{\Delta}/C_{\min}$ and simplify our lower bound to

$$\begin{aligned} |\hat{z}_i| \geq \|[Q^*_{S^cS}{Q^*_{SS}}^{-1}\hat{z}_S]_i\|_\infty - \|T_{1,i}\|_1 - \|T_{2,i}\|_1 - \|T_{3,i}\|_1 \\ -\Big|\frac{W^n_i}{\lambda}\Big| - \Big|\frac{R^n_i}{\lambda}\Big| - \frac{2\sqrt{\Delta}}{C_{\min}}\left(\Big\|\frac{W^n_S}{\lambda}\Big\|_\infty + \Big\|\frac{R^n_S}{\lambda}\Big\|_\infty\right). \end{aligned} \tag{44}$$

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

Taking into account that $\sigma_{\min}(Q^*_{SS}) \leq \max_{ij} Q^*_{ij} \leq 1$ and that $\Delta > 1$, the last condition on $\mathcal{E}_i$ immediately bounds all terms involving W^n by $\epsilon/8$. Some straightforward manipulations imply (see Lemma 7 from [11] for a similar computation)

$$\begin{aligned} &\|T_{1,i}\|_1 \leq \frac{\Delta}{C_{\min}^2}\|Q^n_{SS} - Q^*_{SS}\|_\infty, \qquad \|T_{2,i}\|_1 \leq \frac{\sqrt{\Delta}}{C_{\min}}\|[Q^n_{S^cS} - Q^*_{S^cS}]_i\|_\infty, \\ &\|T_{3,i}\|_1 \leq \frac{2\Delta}{C_{\min}^2}\|Q^n_{SS} - Q^*_{SS}\|_\infty\|[Q^n_{S^cS} - Q^*_{S^cS}]_i\|_\infty, \end{aligned}$$

and thus, again making use of the fact that $\sigma_{\min}(Q^*_{SS}) \leq 1$, all will be bounded by $\epsilon/8$ when $\mathcal{E}_i$ holds. The final step of the proof consists in showing that if $\mathcal{E}_i$ holds and λ satisfies the condition given in the Lemma enunciation then the terms involving R^n will also be bounded above by $\epsilon/8$. The details of this calculation are included in Appendix C.1. □

Proof. (Lemma 4.3.) Let us state explicitly the local weak convergence result mentioned in Sec. 4.3 right before our statement of Lemma 4.3. For $t \in \mathbb{N}$, let $\mathsf{T}(t) = (V_\mathsf{T}, E_\mathsf{T})$ be the regular rooted tree of degree Δ of t generations and define the associated Ising measure as

$$\mu^+_{\mathsf{T},\theta}(\underline{x}) = \frac{1}{Z_{\mathsf{T},\theta}} \prod_{(i,j)\in E_\mathsf{T}} e^{\theta x_i x_j} \prod_{i\in\partial\mathsf{T}(t)} e^{h^* x_i}. \tag{45}$$

Here $\partial\mathsf{T}(t)$ is the set of leaves of $\mathsf{T}(t)$ and h^* is the unique positive solution of

$$h = (\Delta-1)\,\mathrm{atanh}\left\{\tanh\theta\,\tanh h\right\}. \tag{46}$$

It was proved in [33] 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 (45))

$$\lim_{p\to\infty} \mathbb{E}_{G,\theta}\{g(\underline{X}_{\mathsf{B}_r(t)})\} = \mathbb{E}_{\mathsf{T}(t),\theta,+}\{g(\underline{X}_{\mathsf{T}(t)})\}\,. \tag{47}$$

Notice that this characterizes expectations completely since if $g(\underline{x}_{\mathsf{B}_r(t)}) = -g(-\underline{x}_{\mathsf{B}_r(t)})$ then,

$$\mathbb{E}_{G,\theta}\{g(\underline{X}_{\mathsf{B}_r(t)})\} = 0. \tag{48}$$

The proof consists in considering $[Q^*_{S^cS}{Q^*_{SS}}^{-1}\hat{z}^*_S]_i$ for $t = \mathrm{dist}(r,i)$ bounded. We then write $(Q^*_{SS})_{lk} = \mathbb{E}_{G,\theta}\{g_{l,k}(\underline{X}_{\mathsf{B}_r(t)})\}$ and $(Q^*_{S^cS})_{il} = \mathbb{E}_{G,\theta}\{g_{i,l}(\underline{X}_{\mathsf{B}_r(t)})\}$ for some functions $g_{\cdot,\cdot}(\underline{X}_{\mathsf{B}_r(t)})$ and apply the weak convergence result (47) to these expectations. We thus reduced the calculation of $[Q^*_{S^cS}{Q^*_{SS}}^{-1}\hat{z}^*_S]_i$ to the calculation of expectations with respect to the tree measure (45). 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 is deferred to Appendix C.3.

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

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.

A Covariance calculation for the toy example

In this section we compute the covariance matrix for the Ising model on the graph G_p introduced within the toy example of Section 1.1, see Fig. 1. In fact we only need to compute $\mathbb{E}_{G_p,\theta}\{x_1x_2\}$, $\mathbb{E}_{G_p,\theta}\{x_1x_3\}$ and $\mathbb{E}_{G_p,\theta}\{x_3x_4\}$, since all other covariances reduce to one of these tree by symmetry. First recall that by [35] we can write the correlation between x_i and x_j as follows

$$\mathbb{E}_{G,\theta}\{x_ix_j\} = \frac{\sum_{F\in\mathcal{I}(G)}(\tanh\theta)^{|F|}}{\sum_{G\in\mathcal{P}(G)}(\tanh\theta)^{|F|}}, \tag{49}$$

where:(i) $\mathcal{I}(G)$ is the set of all subsets of edges of graphs of G with odd number of edges adjacent to node i and j and even number of edges adjacent to every other node; (ii) $\mathcal{P}(G)$ is the set of all subsets of edges of G with even number of edges in all nodes; (iii) $|F|$ is the number of edges in F.

Expression (49) implies three basic facts that we will use to compute the correlations of G_p. Some of these observations can be proved in different and maybe simpler ways but for a matter of unity, we will explain them from the point of view of (49).

First, if i, j are two nodes in a graph G and k, l two nodes in a graph G' and we 'glue' j and k together (i.e. we fix $x_j = x_k$) to form a new graph G'' (see Figure 4 (a)) then

$$\mathbb{E}_{G'',\theta}\{x_ix_l\} = \mathbb{E}_{G,\theta}\{x_ix_j\}\,\mathbb{E}_{G',\theta}\{x_kx_l\}. \tag{50}$$

Second, if instead we 'glue' i with k and j with l (i.e. we fix $x_i = x_k$ and $x_j = x_l$) (see Figure 4 (b)) then

$$\mathbb{E}_{G'',\theta}\{x_ix_j\} = \frac{\mathbb{E}_{G,\theta}\{x_ix_j\} + \mathbb{E}_{G',\theta}\{x_kx_l\}}{1 + \mathbb{E}_{G,\theta}\{x_ix_j\}\,\mathbb{E}_{G',\theta}\{x_kx_l\}} \tag{51}$$

$$= \tanh(\operatorname{arctanh}(\mathbb{E}_{G,\theta}\{x_ix_j\}) + \operatorname{arctanh}(\mathbb{E}_{G',\theta}\{x_kx_l\})). \tag{52}$$

Note that in this second case we are computing $\mathbb{E}_{G'',\theta}\{x_ix_j\}$ and not $\mathbb{E}_{G'',\theta}\{x_ix_l\}$.

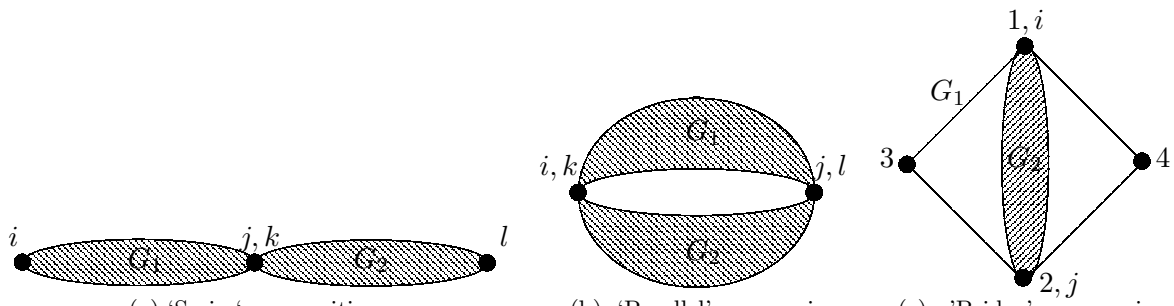
Finally, if G is the square graph formed by nodes $\{1, 2, 3, 4\}$ and edge set $\{(1, 3), (1, 4), (2, 3), (2, 4)\}$ and G' is some other graph to which nodes i and j belong and we 'glue' node 1 with i and node 2 with j (i.e. $x_1 = x_i$ and $x_2 = x_j$) to form G'' (see Figure 4 (c)) then

$$\mathbb{E}_{G_{12},\theta}\{x_3x_4\} = \frac{2\tanh^2\theta + 2\mathbb{E}_{G_2,\theta}\{x_ix_j\}\tanh^2\theta}{1 + \tanh^4\theta + 2\mathbb{E}_{G_2,\theta}\{x_ix_j\}\tanh^2\theta}. \tag{53}$$

With these three relationships we can quickly compute $\mathbb{E}_{G_p,\theta}\{x_1x_2\}$, $\mathbb{E}_{G_p,\theta}\{x_1x_3\}$ and $\mathbb{E}_{G_p,\theta}\{x_3x_4\}$. Let $p = 3$ and note that from (50) we have that $\mathbb{E}_{G_3,\theta}\{x_1x_2\} = \tanh^2\theta$. Since G_p is formed by $p-2$ copies of G_3 glued in 'parallel' in between nodes 1 and 2, by (52) we have that $\mathbb{E}_{G_p,\theta}\{x_1x_2\} = \tanh((p-2)\operatorname{arctanh}(\tanh^2\theta))$. Now notice that G_p can also be seen as a single edge connecting 1 and 3 in 'parallel' with the graph formed by connecting in 'series' the edge $(2, 3)$ to a copy of G_{p-1}. This tells us that $\mathbb{E}_{G_p,\theta}\{x_1x_3\} = \tanh(\theta + \operatorname{arctanh}(\mathbb{E}_{G_{p-1},\theta}\{x_1x_2\}\tanh\theta))$. Finally, we can also see G_p as a square graph formed by nodes $\{1, 2, 3, 4\}$ and edges $\{(1, 3), (1, 4), (2, 3), (2, 4)\}$ to which we add G_{p-2} as a 'bridge' in between nodes 1 and 2. Making use of (53) we get that

$$\mathbb{E}_{G_p,\theta}\{x_3x_4\} = \frac{2\tanh^2\theta + 2\mathbb{E}_{G_{p-2},\theta}\{x_1x_2\}\tanh^2\theta}{1 + \tanh^4\theta + 2\mathbb{E}_{G_{p-2},\theta}\{x_1x_2\}\tanh^2\theta}. \tag{54}$$

From these closed form expressions it is now easy to obtain the behavior of the correlations for the regime $\theta \gg 1$ and $p \ll 1$.



(a) 'Series' composition (b) 'Parallel' composition (c) 'Bridge' composition

Figure 4: Correlation for different composite graphs

B Success of regularized logistic regression for small θ

Proof. (Theorem 2.6) The proof of this theorem consists in verifying the conditions of Theorem 1 in [11] (denoted there by A1 and A2) and computing the probability of success as a function of n with slightly more care.

In what follows, $C_{\min}$ is a lower bound for $\sigma_{\min}(Q^*_{SS})$ and $D_{\max}$ [5] is an upper bound for

$$\sigma_{\max}(\mathbb{E}_{G,\underline{\theta}^*}(X_S X_S^T)). \tag{55}$$

We define $1-\alpha \equiv \|Q^*_{S^cS}{Q^*_{SS}}^{-1}\|_\infty$ and $\theta_{\min}$ is the minimum absolute value of the components of $\underline{\theta}^*$. Throughout this proof we will have $\hat{C}_{\min}$ and $\hat{D}_{\max}$ denote $\sigma_{\min}({Q^n}^*_{SS})$ and $\sigma_{\max}\left(\frac{1}{n}\sum_{l=1}^n x_S^{(l)} {x^{(l)}}_S^T\right)$ respectively and $1-\hat{\alpha} = \|Q^n_{S^cS}{Q^n_{SS}}^{-1}\|_\infty$.

Consider the event, $\mathcal{E}$, that the following conditions hold (these conditions are part of the conditions required for Theorem 1 in [11] to be applicable and are labeled by the names of the theorems that use them that help proving Theorem 1),

$$\text{In Lemma 5:} \quad \|Q^n_{SS} - Q^*_{SS}\|_2 < C_{\min}/2\,, \tag{56}$$

$$\text{In Lemma 6: for T1} \quad \sigma_{\min}({Q^n}^*_{SS}) \le C_{\min}/2\,, \tag{57}$$

$$\text{for T1} \quad \|Q^n_{SS} - Q^*_{SS}\|_\infty \le \frac{1}{12}\frac{\alpha}{1-\alpha} C_{\min}/\sqrt{\Delta}\,, \tag{58}$$

$$\text{for T2} \quad \|Q^n_{S^cS} - Q^*_{S^cS}\|_\infty \le \frac{\alpha}{6} C_{\min}/\sqrt{\Delta}\,, \tag{59}$$

$$\text{for T3} \quad \|Q^n_{S^cS} - Q^*_{S^cS}\|_\infty \le \sqrt{\alpha/6}\,, \tag{60}$$

$$\text{In Lemma 7:} \quad \sigma_{\min}({Q^n}^*_{SS}) \le C_{\min}/2 \text{ and } \|Q^n_{SS} - Q^*_{SS}\|_2 \le \sqrt{\frac{\alpha}{6}}\frac{C^2_{\min}}{2\sqrt{\Delta}}\,, \tag{61}$$

[5] It is easy to prove that $C_{\min} \le D_{\max}$

$$\text{In Proposition 1:} \qquad \frac{\|W^n\|_\infty}{\lambda} < \frac{\hat{\alpha}}{4(2-\hat{\alpha})}\,, \tag{62}$$

$$\lambda\Delta \leq \frac{\hat{C}^2_{\min}}{10\hat{D}_{\max}}\,, \tag{63}$$

$$\frac{5}{\hat{C}_{\min}}\lambda\sqrt{\Delta} \leq \frac{\theta_{\min}}{2}\,. \tag{64}$$

Note that these conditions imply that $\hat{C}_{\min} \geq C_{\min}/2$, $\hat{D}_{\max} \leq 2D_{\max}$ and also, from the proof of Proposition 1 in [11], they imply that without loss of generality we can assume $\hat{\alpha} = \alpha/2$. Since the assumption $\sigma_{\min}(Q^{n*}_{SS}) \leq C_{\min}/2$ follows from the assumption of Lemma 5 in [11], all assumptions are in fact assumptions on the proximity, under different norms, of empirical vectors and matrices to their correspondent mean values.

Having the definition of $\mathcal{E}$ in mind we beging by noting that Theorem 1 can be rewritten in the following form.

Theorem B.1. *If $\lambda \leq (C_{\min}/2)^2/(20\Delta D_{\max})$, $\lambda \leq (C_{\min}/2)\theta_{\min}/(20\sqrt{\Delta})$, $1-\alpha < 1$ and $\mathcal{E}$ holds then* RIr *will not fail.*

A straightforward application of Azuma's inequality yields the following upper bound on the probability of these assumptions not occurring together, (the first three terms are for the conditions involving matrix Q and the fourth with the event dealing with matrix W),

$$\begin{aligned}\mathbb{P}_{n,G,\theta}(\mathcal{E}^c) \leq\ & 2e^{-n\frac{1}{32\Delta^2}(d^{(2)}_{SS})^2+2\log\Delta} + 2e^{-n\frac{1}{32\Delta^2}(d^{(\infty)}_{SS})^2+2\log\Delta} \\ & + 2e^{-n\frac{1}{32\Delta^2}(d^{(\infty)}_{S^cS})^2+\log\Delta+\log p-\Delta} + 2e^{-n\frac{\lambda^2}{2^7}(\frac{\hat{\alpha}}{2-\hat{\alpha}})^2+\log p}\,,\end{aligned} \tag{65}$$

where

$$d^{(2)}_{SS} = \min\Big\{\frac{C_{\min}}{2}, \sqrt{\frac{\alpha}{6}}\frac{C^2_{\min}}{2\sqrt{\Delta}}\Big\}, \tag{66}$$

$$d^{(\infty)}_{SS} = \frac{1}{12}\frac{\alpha}{1-\alpha}\frac{C_{\min}}{\sqrt{\Delta}}, \tag{67}$$

$$d^{(\infty)}_{S^cS} = \min\Big\{\frac{\alpha}{6}\frac{C_{\min}}{\sqrt{\Delta}}, \sqrt{\frac{\alpha}{6}}\Big\}. \tag{68}$$

Under the assumption that $\theta \leq K_1/\Delta$ for K_1 small enough we now calculate lower bounds for $C_{\min}$ and α and upper bound for $D_{\max}$ which will allow us to verify the condition of Theorem B.1 and simplify expression for the upper bound on $\mathbb{P}_{n,G,\theta}(\mathcal{E}^c)$.

First notice that by (22) we have $C_{\min} = \sigma_{\min}\{\mathbb{E}_{G,\underline{\theta}^*}((1-\tanh^2\theta M)X_SX_S^T)\}$ where $M = \sum_{t\in\partial r} X_t$. Since $\theta M \leq \theta\Delta \leq K_1$ and because $\sigma_{\min}(AB) \geq \sigma_{\min}(A)\sigma_{\min}(B)$ we have, $C_{\min} \geq (1-K_1^2)\sigma_{\min}(\mathbb{E}_{G,\underline{\theta}^*}\{X_SX_S^T\})$. Now write $\mathbb{E}_{G,\underline{\theta}^*}\{X_SX_S^T\} \equiv \mathbb{1} + Q$ and notice that by (29) Q is a symmetric matrix whose values are non-negative and smaller than $\tanh\theta/(1-\Delta\tanh\theta)$. Since $\sigma_{\min}(\mathbb{E}_{G,\underline{\theta}^*}\{X_SX_S^T\}) = 1 - v^T(-Q)v$ for some unit norm vector v and since, by Cauchy–Schwarz inequality, we have $v^T(-Q)v \leq \|v\|_1^2 \max_{ij}|Q_{ij}| \leq \Delta\tanh\theta/(1-\Delta\tanh\theta) \leq K_1/(1-K_1)$, it follows that $\sigma_{\min}(\mathbb{E}_{G,\underline{\theta}^*}\{X_SX_S^T\}) \geq (1-2K_1)/(1-K_1)$. Consequently, $C_{\min} \geq (1+K_1)(1-2K_1)$. With the bound (29), and again for K_1 small, we can write $D_{\max} \leq 1+\Delta\tanh\theta/(1-\Delta\tanh\theta) \leq (1-K_1)^{-1}$. A similar calculation yields $1-\alpha \leq K_1/((1-K_1^2)(1-2K_1))$.

For K_1 small enough, and looking at the bounds just obtained for $C_{\min}$ and $D_{\min}$, the restriction on λ in Theorem B.1, namely

$$\lambda \leq C_{\min}/40\sqrt{\Delta}\min\{\theta, C_{\min}/40D_{\max}\sqrt{\Delta}\}, \tag{69}$$

can be simplified to $C_{\min}\theta/40\sqrt{\Delta}$. Choosing $\lambda = K_3\theta\Delta^{-1/2}$ it is easy to see we can simplify the expression for the probability upper bound and write

$$1 - \mathbb{P}_{n,G,\theta}(\mathcal{E}) \leq 8e^{-nK_2^{-1}\theta^2\Delta^{-1}+2\log p} \tag{70}$$

for some constant K_2 which in turn implies the bound on $n_{\mathsf{Rlr}}(\lambda)$. □

C Failure of regularized logistic regression at large θ

C.1 Bound on terms involving R^n

Proof. (Lemma 4.1) We outline here the upper bound on the term R^n. Note that we are omitting the samples $\{x^{(i)}\}$ in the argument of function L and we are representing $\underline{\theta}_{r,\cdot}$ by $\underline{\theta}$. This proof is just a replica and fusion of Lemmas 3 and 4 in [11]. Through out this proof we have $\hat{\underline{\theta}}_{S^C} = \underline{\theta}^*_{S^C} = 0$.

First we write,

$$R^n_j = [\nabla^2 L(\bar{\underline{\theta}}^{(j)}) - \nabla^2 L(\underline{\theta}^*)]^T_j [\hat{\underline{\theta}} - \underline{\theta}^*] \tag{71}$$

$$= \frac{1}{n}\sum_{i=1}^{n}[\eta(\bar{\underline{\theta}}^{(j)}) - \eta(\underline{\theta}^*)][x^{(i)}\, x^{(i)^T}]^T_j [\hat{\underline{\theta}} - \underline{\theta}^*] \tag{72}$$

for some point $\bar{\underline{\theta}}^{(j)}$ lying in the line between $\hat{\underline{\theta}}$ and $\underline{\theta}^*$,i.e. $\bar{\underline{\theta}}^{(j)} = t_j\hat{\underline{\theta}} + (1-t_j)\underline{\theta}^*$. Since $\eta(\underline{\theta}) = g(x_r^{(i)}\sum_{t\in V\setminus r}\theta_{rt}x_t^{(i)}) = g(x_r^{(i)}\underline{\theta}^T\underline{x}^{(i)}) = g(\underline{\theta}^T\underline{x}^{(i)})$ where $g(s) = 4e^{2s}/(1+e^{2s})^2$ another application of the chain rule yields,

$$R^n_j = \frac{1}{n}\sum_{i=1}^{n} g'(\bar{\bar{\underline{\theta}}}^{(j)^T}x^{(i)})x^{(i)^T}[\bar{\underline{\theta}}^{(j)} - \underline{\theta}^*]\{x_j^{(i)}\, x^{(i)^T}[\hat{\underline{\theta}} - \underline{\theta}^*]\} \tag{73}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\{g'(\bar{\bar{\underline{\theta}}}^{(j)^T}x^{(i)})x_j^{(i)}\}\{[\bar{\underline{\theta}}^{(j)} - \underline{\theta}^*]^T x^{(i)}\, x^{(i)^T}[\hat{\underline{\theta}} - \underline{\theta}^*]\} \tag{74}$$

where $\bar{\bar{\underline{\theta}}}^{(j)}$ is a point in the line between $\bar{\underline{\theta}}^{(j)}$ and $\underline{\theta}^*$. Let

$$b_i := [\bar{\underline{\theta}}^{(j)} - \underline{\theta}^*]^T x^{(i)}\, x^{(i)^T}[\hat{\underline{\theta}} - \underline{\theta}^*] = t_j[\hat{\underline{\theta}} - \underline{\theta}^*]^T x^{(i)}\, x^{(i)^T}[\hat{\underline{\theta}} - \underline{\theta}^*] \geq 0 \tag{75}$$

then, noticing that $\hat{\underline{\theta}}_{S^C} = \underline{\theta}^*_{S^C} = 0$ and $|g'| \leq 1$ we can apply Holder's inequality to obtain,

$$|R^n_j| \leq \frac{1}{n}\|b\|_1 \leq t_j[\hat{\underline{\theta}}_S - \underline{\theta}^*_S]^T\left\{\frac{1}{n}\sum_{i=1}^{n} x_S^{(i)}\, x_S^{(i)^T}\right\}[\hat{\underline{\theta}}_S - \underline{\theta}^*_S] \leq \Delta\|\hat{\underline{\theta}}_S - \underline{\theta}^*_S\|_2^2. \tag{76}$$

Slightly readapting the proof of Lemma 3 from [11] we now show that

$$\|\hat{\underline{\theta}}_S - \underline{\theta}^*_S\|_2 \leq \frac{C_{\min}}{4\Delta^{3/2}}\left(1 - \sqrt{1 - \lambda\frac{16\Delta^2}{C_{\min}^2}\left(1 + \left\|\frac{W_S^n}{\lambda}\right\|_\infty\right)}\right). \tag{77}$$

Define $G(u) = L(\underline{\theta}^* + u) - L(\underline{\theta}^*) + \lambda(\|\underline{\theta}^* + u\|_1 - \|\underline{\theta}^*\|_1)$. Since $G(0) = 0$ and G is strictly convex we have that if $G(u) > 0$ for $\|u\|_2 = B$ then $\|\hat{u}\|_2 < B$, where $\hat{u} = \hat{\underline{\theta}} - \underline{\theta}^*$ is the unique minimum point of $G(u)$. To prove (77) we will compute a lower bound on the set of points for which $G(u) > 0$.

By the mean value theorem we can write,

$$G(u) = -W^{nT}u + u^T\nabla^2 L(\underline{\theta}^* + \alpha u)u + \lambda(\|\underline{\theta}^* + u\|_1 - \|\underline{\theta}^*\|_1). \tag{78}$$

Note that $W^n = -\nabla L(\underline{\theta}^*)$.

We now get bounds on each of the terms of the previous expression,

$$\begin{aligned}
|W^{nT}u| &\leq \|W_S^n\|_\infty \sqrt{\Delta}\|u\|_2, && (79)\\
\lambda(\|\underline{\theta}^* + u\|_1 - \|\underline{\theta}^*\|_1) &\geq -\lambda\sqrt{\Delta}\|u\|_2 && (80)\\
u^T\nabla^2 L(\underline{\theta}^* + \alpha u)u &\geq \|u\|_2^2 \left(\sigma_{\min}(Q_{SS}^n{}^*) - \Delta^{1/2}\|u\|_2\, \sigma_{\max}\left(\frac{1}{n}\sum_{i=1}^n x_S^{(i)} x_S^{(i)T}\right)\right), \\
&\geq \|u\|_2^2 \left(C_{\min}/2 - \Delta^{3/2}\|u\|_2\right). && (81)
\end{aligned}$$

Thus can write,

$$G(u) \geq \|u\|_2\, \Delta^{1/2}\left(-\Delta\|u\|_2^2 + \Delta^{-1/2}\frac{C_{\min}}{2}\|u\|_2 - \lambda - \|W_S^n\|_\infty\right), \tag{82}$$

from which we derive expression (77).

If $\mathcal{E}_i$ holds we can assume without loss of generality $\|\frac{W_S^n}{\lambda}\|_\infty < \epsilon$. Now notice that $1 - \sqrt{1-x} \leq x, x \in [0,1]$ and thus we can write,

$$|R_j^n| \leq \Delta\left(\frac{C_{\min}}{4\Delta^{3/2}}\frac{16(1+\epsilon)\Delta^2\lambda}{C_{\min}^2}\right)^2 \leq \frac{16\Delta^2\lambda^2(1+\epsilon)^2}{C_{\min}^2}. \tag{83}$$

If we now want that

$$\frac{\Delta|R_j^n|}{\lambda C_{\min}} \leq \frac{\epsilon}{8}, \tag{84}$$

then we can simply impose that $\lambda < C_{\min}^3\epsilon/(2^7(1+\epsilon^2)\Delta^3)$, which finishes the proof.

□

C.2 $n\lambda^2$ must be unbounded with p

Proof. (Lemma 4.2) In this proof $S = \{i\}$ and $S^C = \partial r \backslash \{i\}$.

We prove the lemma by computing a lower bound on the probability that $\|\nabla_{S^C} L(\hat{\underline{\theta}}; \{\underline{x}^{(\ell)}\})\|_\infty > \lambda$ under the assumption that $n\lambda^2 \leq K$ and $\hat{\underline{\theta}}_{S^C} = 0$ and $\hat{\underline{\theta}}_S > 0$ [6]. This will prove the corresponding upper bound on the probability of success of $\mathsf{Rlr}(\lambda)$.

First we show that there exists an $C(\theta)$ such that if $\|\hat{\underline{\theta}}_S\|_\infty > C$ then with high probability $\mathsf{Rlr}(\lambda)$ fails.

Begin by noticing that $\mathbb{E}_{G,\theta}(L(\hat{\underline{\theta}})) \geq \hat{\theta}_{ir}(1 - \tanh\theta)$ and that $|L(\hat{\underline{\theta}}) - \mathbb{E}_{G,\theta}(L(\hat{\underline{\theta}}))| \leq 2\log 2 + 2\|\hat{\underline{\theta}}\|_1$. Then use Azuma's inequality to get the following bound,

$$\begin{aligned}
&\mathbb{P}_{n,G,\theta}(L(\hat{\underline{\theta}}) + \lambda\|\hat{\underline{\theta}}\|_1 > L(\underline{0})) && (85)\\
&= \mathbb{P}_{n,G,\theta}(L(\hat{\underline{\theta}}) - \mathbb{E}_{G,\theta}(L(\hat{\underline{\theta}})) > \log 2 - \lambda\|\hat{\underline{\theta}}\|_1 - \mathbb{E}_{G,\theta}(L(\hat{\underline{\theta}}))) && (86)\\
&\geq 1 - e^{\frac{-2n(\log 2 - \lambda\hat{\theta}_{ir} - \mathbb{E}_{G,\theta}(L(\hat{\underline{\theta}})))^2}{(2\log 2 + 2\hat{\theta}_{ir})^2}}. && (87)
\end{aligned}$$

[6] The requirement $\hat{\theta}_{ri} > 0$, necessary for correct reconstruction, allows us to ignore the $\|.\|_\infty$ and $\|.\|_1$ in what follows.

If $\|\underline{\hat{\theta}}_S\|_\infty > C(\theta)$ for large enough $C(\theta)$ then we can lower bound the previous expression by

$$1 - e^{\frac{-2n(\frac{\hat{\theta}_{ir}}{2})^2(1-\tanh\theta)^2}{(2\log 2+2\hat{\theta}_{ir})^2}} \geq 1 - e^{\frac{-nC^2(1-\tanh\theta)^2}{8(\log 2+C)^2}} \geq 1 - e^{\frac{-n(1-\tanh\theta)^2}{32}}. \tag{88}$$

Since $L(\underline{\hat{\theta}}) + \lambda\|\underline{\hat{\theta}}\|_1 > L(\underline{0})$ contradicts the fact that $\underline{\hat{\theta}}$ is the optimal solution found by Rlr this shows that with high probability $\hat{\theta}_{ir}$ must be smaller than $C(\theta)$.

Under the assumption that $\hat{\theta}_{ri} \leq C$ and $n\lambda^2 \leq K$ we will now compute a lower bound for the event $\|\nabla_{S^C} L(\underline{\hat{\theta}})\|_\infty > \lambda$.

$$\mathbb{P}_{n,G,\theta}\left(\|\frac{1}{\lambda}\nabla_{S^C} L(\underline{\hat{\theta}})\|_\infty > 1 \middle| \hat{\theta}_{ri} \leq C\right) \tag{89}$$

$$= \mathbb{P}_{n,G,\theta}\left(\|\frac{1}{\lambda}\frac{1}{n}\sum_{\ell=1}^{n} X_r^{(\ell)} X_{S^C}^{(\ell)}(1-\tanh(X_r^{(\ell)}X_i^{(\ell)}\hat{\theta}_{ri}))\|_\infty > 1\right) \tag{90}$$

$$\geq 1 - \mathbb{P}_{n,G,\theta}\left(\frac{1}{\sqrt{n}}\sum_{\ell=1}^{n} X_r^{(\ell)} X_{r'}^{(\ell)}(1-\tanh(X_r^{(\ell)}X_i^{(\ell)}\hat{\theta}_{ri})) \leq \sqrt{K}\ , \forall r' \in S^C\right) \tag{91}$$

$$\geq 1 - \mathbb{E}_{G,\theta}\left(\mathbb{P}_{n,G,\theta}\left(\frac{1}{\sqrt{n}}\sum_{\ell=1}^{n} X_{r'}^{(\ell)} C_\ell \leq \sqrt{K}\ , \forall r' \in S^C \middle| \{C_\ell\}_{\ell=1}^n\right)\right), \tag{92}$$

where $C_\ell = X_r^{(\ell)}(1-\tanh(X_r^{(\ell)}X_i^{(\ell)}\hat{\theta}_{ri}))$.

Conditioned on $\{C_\ell\}_{l=1}^n$ all the $\sum_\ell X_{r'}^{(\ell)} C_\ell$ are independent and identically distributed. Hence, choosing one particular $r_0' \in S_C$, and defining $V_\ell = X_{r_0'}^{\ell}$ we can rewrite the previous expression as,

$$1 - \mathbb{E}_{G,\theta}\left(\mathbb{P}_{n,G,\theta}\left(\frac{1}{\sqrt{n}}\sum_{\ell=1}^{n} V_\ell C_\ell \leq \sqrt{K} \middle| \{C_\ell\}_{\ell=1}^n\right)^{p-1}\right). \tag{93}$$

We now use the central limit theorem for independent nonidentical random variables to upper bound the conditional probability inside the expectation. It is easy to see that Lyapunov conditions hold. In fact, let $s_n^2 = \sum_{\ell=1}^n \mathrm{Var}(V_\ell C_\ell | \{C_\ell\}_{\ell=1}^n) = \sum_{\ell=1}^n C_\ell^2$ then for some $\delta > 0$,

$$\mathbb{E}_{G,\theta}(|V_\ell C_\ell|^{2+\delta} | \{C_\ell\}_{\ell=1}^n) = |C_\ell|^{2+\delta} < \infty\ , \forall \ell \tag{94}$$

$$\tag{95}$$

and

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{\ell=1}^{n} \mathbb{E}_{G,\theta}(|V_\ell C_\ell - \mathbb{E}_{G,\theta}(V_\ell C_\ell | \{C_\ell\}_{\ell=1}^n)|^{2+\delta} | \{C_\ell\}_{\ell=1}^n) \tag{96}$$

$$= \lim_{n\to\infty} \frac{1}{(\sum_{\ell=1}^n C_\ell^2)^{1+\delta/2}} \sum_{\ell=1}^{n} |C_\ell|^{2+\delta} \leq \lim_{n\to\infty} n^{-\delta/2}\left(\frac{1+\tanh C(\theta)}{1-\tanh C(\theta)}\right)^{2+\delta} = 0. \tag{97}$$

Thus we can write,

$$\mathbb{P}_{n,G,\theta}\left(\frac{1}{\sqrt{n}}\sum_{\ell=1}^n V_\ell C_\ell \leq \sqrt{K} \middle| \{C_\ell\}_{\ell=1}^n\right) = \mathbb{P}_{n,G,\theta}\left(\frac{\sum_{\ell=1}^n V_\ell C_\ell}{\sqrt{\sum_{\ell=1}^n C_\ell^2}} \leq \sqrt{K}\frac{\sqrt{n}}{\sqrt{\sum_{\ell=1}^n C_\ell^2}} \middle| \{C_\ell\}_{\ell=1}^n\right) \tag{98}$$

$$\leq \mathbb{P}_{n,G,\theta}\left(\frac{\sum_{\ell=1}^n V_\ell C_\ell}{\sqrt{\sum_{\ell=1}^n C_\ell^2}} \leq \frac{\sqrt{K}}{1-\tanh C(\theta)} \middle| \{C_\ell\}_{\ell=1}^n\right) = \Phi\left(\frac{\sqrt{K}}{1-\tanh C(\theta)}\right) + \epsilon_n \tag{99}$$

where Φ is the cumulative distribution of the normal(0,1) distribution and $\epsilon_n \to 0$ with n. We can finally write,

$$\mathbb{P}_{n,G,\theta}\left(\|\frac{1}{\lambda}\nabla_{S^C}L(\underline{\hat{\theta}})\|_\infty > 1\middle|\hat{\theta}_{ri} \le C\right) \ge 1 - e^{(p-1)(\log(\Phi\left(\frac{\sqrt{K}}{1-\tanh C(\theta)}\right)+\epsilon_n))} \ge 1 - e^{-pM(K,\theta)} \tag{100}$$

for n big enough. In the above expression $M(K,\theta) \to 0$ as $K \to \infty$. From this bound and (88) we get the desired upper bound on the probability of success of RIr. □

C.3 Random regular graphs and the violation of the incoherence condition

Proof. (Lemma 4.3) We explain here the calculations with respect to the tree model (45). Throughout all calculations we assume that $0 < \theta < \infty$. An important property that follows from the fixed point equation (46) is that, if $g(\underline{x}_{\mathsf{T}(t)})$ is a function of the variables in $\mathsf{T}(t)$ then

$$\mathbb{E}_{\mathsf{T}(t),\theta,+}\{g(\underline{X}_{\mathsf{T}(t)})\} = \mathbb{E}_{\mathsf{T}(t+1),\theta,+}\{g(\underline{X}_{\mathsf{T}(t)})\}\ , \tag{101}$$

with the obvious identification of $\mathsf{T}(t)$ as a subtree of $\mathsf{T}(t+1)$.

Let r be a uniformly random vertex in G and $i \neq j$ two neighbors of r. Using the local weak convergence property (47) with $t = 1$ we get

$$\lim_{p\to\infty}(Q^*_{SS})_{ii} \equiv a = \mathbb{E}_{\mathsf{T}(1),\theta,+}\Big(\frac{1}{\cosh^2\theta M}\Big)\,, \tag{102}$$

$$\lim_{p\to\infty}(Q^*_{SS})_{ij} \equiv b = \mathbb{E}_{\mathsf{T}(1),\theta,+}\Big(\frac{X_iX_j}{\cosh^2\theta M}\Big), \tag{103}$$

where $M \equiv \sum_{i\in\partial\mathsf{T}(1)} X_i$ is the sum of the variables on the leaves of a depth 1 tree, and $i, j \in \partial\mathsf{T}(1)$. For r' at distance $t > 1$ from r, consider the Δ-dimensional vector in

$$\lim_{p\to\infty}(Q^*_{S^cS})_{r'} = F_S(t)\,. \tag{104}$$

Elements of $F_S(t)$ are of the form $\mathbb{E}_{\mathsf{T}(t),\theta,+}\left(\frac{X_{r'}X_i}{\cosh^2\theta M}\right)$ where $i \in \partial\mathsf{T}(1)$. These elements can take only two different values: one if r' is a child of j and other if not. We denote the first value by $\mathrm{F_d(t)}$ and the second by $\mathrm{F_i(t)}$. Since $\hat{z}^*_S = \mathbb{1}$ is an eigenvector of Q^*_{SS} with eigenvalue $a+(\Delta-1)b$ we can write,

$$\lim_{p\to\infty}\|Q^*_{S^CS}Q^{*-1}_{SS}\hat{z}^*_S\|_\infty = \sup_{t\ge1}|A(t)| \tag{105}$$

where

$$A(t) = \frac{F_d(t)+(\Delta-1)F_i(t)}{a+(\Delta-1)b} = \frac{\mathbb{E}_+(X_{r'}M/\cosh^2(\theta M))}{\mathbb{E}_+(X_iM/\cosh^2(\theta M))}. \tag{106}$$

In this expression, and through the rest of the proof, $\mathbb{E}_+$ will denote $\mathbb{E}_{\mathsf{T}(t'),\theta,+}$ where t' is the smallest value such that all the variables inside the expectation are in $\mathsf{T}(t')$. Now, conditioning on the value of X_i ($i \in \partial\mathsf{T}(1)$) we can write,

$$\mathbb{E}_+(X_{r'}M/\cosh^2(\theta M)) = c_1\langle t\rangle_+ + c_2\langle t\rangle_-, \tag{107}$$

$$\mathbb{E}_+(X_iM/\cosh^2(\theta M)) = c_1 - c_2. \tag{108}$$

where

$$
\begin{aligned}
c_1 &= \mathbb{E}_+(\mathbb{1}_{X_i=1} M/\cosh^2(\theta M)), && (109)\\
c_2 &= \mathbb{E}_+(\mathbb{1}_{X_i=-1} M/\cosh^2(\theta M)), && (110)\\
\langle t\rangle_+ &= \mathbb{E}_+(X_{r'}|X_i=1), && (111)\\
\langle t\rangle_- &= \mathbb{E}_+(X_{r'}|X_i=-1). && (112)\\
& && (113)
\end{aligned}
$$

In the expression above the binomial coefficients are to be assume zero whenever its parameters are not integer values. In order to prove that the incoherence condition is violated we will now show that $B \equiv \lim_{t\to\infty} A(t) > 1$ if θ is large enough. Writing a first order recurrence relation for $\langle t\rangle_+$ and $\langle t\rangle_-$ it is not hard to see that,

$$
\begin{aligned}
\langle t\rangle_+ &= \frac{\beta-\alpha}{\alpha+\beta-2} + \frac{2(\alpha-1)}{\alpha+\beta-2}(\alpha+\beta-1)^t, && (114)\\
\langle t\rangle_- &= \frac{\beta-\alpha}{\alpha+\beta-2} + \frac{2(1-\beta)}{\alpha+\beta-2}(\alpha+\beta-1)^t, && (115)
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha &= \mathbb{P}_+(X_{r''}=1|X_{r'}=1) = e^{h^*+\theta}/(e^{h^*+\theta}+e^{-h^*-\theta}), && (116)\\
\beta &= \mathbb{P}_+(X_{r''}=-1|X_{r'}=-1) = e^{-h^*+\theta}/(e^{h^*-\theta}+e^{-h^*+\theta}), && (117)
\end{aligned}
$$

and r'' denotes a child of r, i.e., a node at distance $t+1$ from r. Recall that h^* is the unique positive solution of (46). In the above expression $\mathbb{P}_+$ denotes the probability associated with the measure (45) where again we can restrict T to the smallest tree containing all the variables that compose the event whose probability we are trying to compute. Since $0 < \alpha+\beta-1 < 1$ we have that

$$
B = \frac{\beta-\alpha}{\alpha+\beta-2}\frac{c_1+c_2}{c_1-c_2}. \tag{118}
$$

A little bit of algebra allows us to write,

$$
\begin{aligned}
\frac{\beta-\alpha}{\alpha+\beta-2} &= \frac{(1-\beta)-(1-\alpha)}{(1-\alpha)+(1-\beta)} && (119)\\
&= \frac{\mathbb{P}_+(X_{r''}=1|X_{r'}=-1)-\mathbb{P}_+(X_{r''}=-1|X_{r'}=1)}{\mathbb{P}_+(X_{r''}=1|X_{r'}=-1)+\mathbb{P}_+(X_{r''}=-1|X_{r'}=1)} && (120)\\
&= \frac{\frac{\mathbb{P}_+(X_{r''}=1,X_{r'}=-1)}{\mathbb{P}_+(X_{r'}=-1)} - \frac{\mathbb{P}_+(X_{r''}=-1,X_{r'}=1)}{\mathbb{P}_+(X_{r'}=1)}}{\frac{\mathbb{P}_+(X_{r''}=1,X_{r'}=-1)}{\mathbb{P}_+(X_{r'}=-1)} + \frac{\mathbb{P}_+(X_{r''}=-1,X_{r'}=1)}{\mathbb{P}_+(X_{r'}=1)}} && (121)\\
&= \frac{1/\mathbb{P}_+(X_{r'}=-1)-1/\mathbb{P}_+(X_{r'}=1)}{1/\mathbb{P}_+(X_{r'}=-1)+1/\mathbb{P}_+(X_{r'}=1)} && (122)\\
&= \frac{\mathbb{P}_+(X_{r'}=1)-\mathbb{P}_+(X_{r'}=-1)}{\mathbb{P}_+(X_{r'}=1)+\mathbb{P}_+(X_{r'}=-1)} = \mathbb{E}_{\mathsf{T}(1),\theta,+}(X_{r'}) = \tanh(\Delta h^*/(\Delta-1)). && (123)
\end{aligned}
$$

In addition, taking into account that c_1 and c_2 can be expressed as,

$$
c_1 = \frac{2}{Z}\sum_{m=-\Delta}^{\Delta}\binom{\Delta-1}{\frac{\Delta+m-2}{2}}\frac{me^{h^*m}}{\cosh\theta m}, \tag{124}
$$

$$
c_2 = \frac{2}{Z}\sum_{m=-\Delta}^{\Delta}\binom{\Delta-1}{\frac{\Delta+m}{2}}\frac{me^{h^*m}}{\cosh\theta m}, \tag{125}
$$

$$
\tag{126}
$$

we have,

$$\frac{c_1+c_2}{c_1-c_2}=\frac{\sum_{m=1}^{\Delta}\binom{\Delta}{\frac{\Delta+m}{2}}m\frac{\sinh mh^*}{\cosh\theta m}}{\sum_{m=1}^{\Delta}\binom{\Delta}{\frac{\Delta+m}{2}}\frac{m^2}{\Delta}\frac{\cosh h^*m}{\cosh\theta m}}. \tag{127}$$

Expanding everything in powers of e^{-h^*} we get,

$$\lim_{p\to\infty}\|Q^*_{S^cS}{Q^*_{SS}}^{-1}\|\geq B=\left(1-2e^{-2h^*\Delta/(\Delta-1)}+...\right)\left(1+(\Delta-2)e^{-2h^*+2}+...\right) \tag{128}$$

$$\left(1-\frac{(\Delta-2)^2}{\Delta}e^{-2h^*+2}+...\right)=1+2\frac{\Delta-2}{\Delta}e^{-2h^*+2}+.... \tag{129}$$

Since h^* grows with θ [7] this expansion proves the first part of Lemma 4.3. In fact, this expression shows that for large θ, as θ increases, B decays to 1 from above. Hence, there exists a $\theta_{\text{thr}}(\Delta)$ such that for all $\theta>\theta_{\text{thr}}(\Delta)$ we will have $\lim_{p\to\infty}\|Q^*_{S^CS}Q^*_{SS}\|_\infty>0$.

Remark C.1. *It is interesting to see that the condition for $B\geq 1$ is equivalent to $c_1(\alpha-1)+c_2(1-\beta)\geq 0$. This implies that if $B\geq 1$ then $A(t)\geq B$ and if $B\leq 1$ then $A(t)\leq B$. Hence, when $B>1$ we have $A(t)>1$ $\forall t$ and when $B<1$ we have $A(t)<1$ $\forall t$. Consequently, $\{\theta: A(t)>1\}=\{\theta: B>1\}$ which does not depend on t. It is not hard to prove that $A(t)>0$ $\forall t,\theta$ and thus, $\{\theta:\lim_{p\to\infty}\|Q^*_{S^CS}{Q^*_{SS}}^{-1}\|_\infty>1\}=\{\theta: B>1\}$.*

We now study how $\theta_{\text{thr}}(\Delta)$ scales with Δ for large Δ. Notice that $B=1$ is equivalent to $S(\theta)\equiv c_1(\alpha-1)+c_2(1-\beta)=0$. It is not hard to see that this equation has a single solution [8]. We show that if we search for solutions, θ, that scale like Δ^{-1} then in the limit when $\Delta\to\infty$ we get an expression that exhibits a single nontrivial zero. This means that for large Δ the solution of $S(\theta)=0$ must be of the form $\tilde\theta\Delta^{-1}(1+o(1))$, where $\tilde\theta$ is the solution of the scaled equation.

First notice that when $\Delta\to\infty$ and $\theta=\tilde\theta/(\Delta-1)$ then h^* converges to the solution of $h^*=\tilde\theta\tanh h^*$. We denote this solution by h^*_∞. Hence, for large finite Δ we can say that $h^*=h^*_\infty+O(\Delta^{-1})$.

We now write new expressions for c_1,c_2,α and β namely,

$$c_1 = \frac{1}{2}\mathbb{E}_+(M/\cosh^2(\theta M))+\frac{1}{2\Delta}\mathbb{E}_+(M^2/\cosh^2(\theta M)), \tag{130}$$

$$c_2 = \frac{1}{2}\mathbb{E}_+(M/\cosh^2(\theta M))-\frac{1}{2\Delta}\mathbb{E}_+(M^2/\cosh^2(\theta M)), \tag{131}$$

$$1-\alpha = \frac{1}{2}(1-\tanh(h^*+\tilde\theta/(\Delta-1))), \tag{132}$$

$$1-\beta = \frac{1}{2}(1+\tanh(h^*-\tilde\theta/(\Delta-1))). \tag{133}$$

$$\tag{134}$$

Expanding the function tanh(.) in α and β in powers of Δ^{-1} we can write

$$S(\theta) = \frac{1}{2}\tanh h^*\,\mathbb{E}_+(M/\cosh^2(\theta M))-\frac{1}{2\Delta}\mathbb{E}_+(M^2/\cosh^2(\theta M)) \tag{135}$$

$$+ \frac{\tilde\theta}{2\Delta(\Delta-1)}\text{sech}^2h^*\,\mathbb{E}_+(M^2/\cosh^2(\theta M))+O(\Delta^{-2}). \tag{136}$$

[7] $h^*=(\Delta-1+o(1))\theta$

[8] By Remark C.1, this tells us that there is a single point where $\lim_{p\to\infty}\|Q^*_{S^CS}{Q^*_{SS}}^{-1}\|_\infty$ crosses 1.

Note that we have not expanded h^* in powers of Δ^{-1}. Defining $\mathbb{E}_+^0(.)$ to be the expectation with respect to the tree model (45) where all connections to node r have been removed (the field on each node is still h^*) we can write,

$$\mathbb{E}_+(M/\cosh^2(\theta M)) = \frac{\mathbb{E}_+^0(M/\cosh(\tilde{\theta}M/(\Delta-1)))}{\mathbb{E}_+^0(\cosh(\tilde{\theta}M/(\Delta-1)))}, \tag{137}$$

$$\mathbb{E}_+(M^2/\cosh^2(\theta M)) = \frac{\mathbb{E}_+^0(M^2/\cosh(\tilde{\theta}M/(\Delta-1)))}{\mathbb{E}_+^0(\cosh(\tilde{\theta}M/(\Delta-1)))}. \tag{138}$$

In addition, making use of the symmetry of the regular tree and expanding $\cosh(\tilde{\theta}M/(\Delta-1))$ around $\tilde{\theta}M'/(\Delta-1)$ and $\tilde{\theta}M''/(\Delta-1)$ (M' and M'' to be defined later) we can write

$$\mathbb{E}_+^0\left(\frac{M}{\cosh(\tilde{\theta}M/(\Delta-1))}\right) = \Delta\,\mathbb{E}_+^0\left(\frac{X_i}{\cosh(\tilde{\theta}M/(\Delta-1))}\right), \tag{139}$$

$$\mathbb{E}_+^0\left(\frac{X_i}{\cosh(\tilde{\theta}M/(\Delta-1))}\right) = \tanh h^*\,\mathbb{E}_+^0\left(\frac{1}{\cosh(\tilde{\theta}M'/(\Delta-1))}\right) \tag{140}$$

$$- \frac{\tilde{\theta}}{\Delta-1}\mathbb{E}_+^0\left(\frac{\tanh(\tilde{\theta}M'/(\Delta-1))}{\cosh(\tilde{\theta}M'/(\Delta-1))}\right) + O(\Delta^{-2}), \tag{141}$$

$$\mathbb{E}_+^0\left(\frac{M^2}{\cosh(\tilde{\theta}M/(\Delta-1))}\right) = \Delta\mathbb{E}_+^0\left(\frac{1}{\cosh(\tilde{\theta}M/(\Delta-1))}\right) \tag{142}$$

$$+ \Delta(\Delta-1)\,\mathbb{E}_+^0\left(\frac{X_iX_j}{\cosh(\tilde{\theta}M/(\Delta-1))}\right), \tag{143}$$

$$\mathbb{E}_+^0\left(\frac{X_iX_j}{\cosh(\tilde{\theta}M/(\Delta-1))}\right) = \tanh^2 h^*\,\mathbb{E}_+^0\left(\frac{1}{\cosh(\tilde{\theta}M''/(\Delta-1))}\right) \tag{144}$$

$$- 2\frac{\tilde{\theta}}{\Delta-1}\tanh h^*\,\mathbb{E}_+^0\left(\frac{\tanh(\tilde{\theta}M''/(\Delta-1))}{\cosh(\tilde{\theta}M''/(\Delta-1))}\right) + O(\Delta^{-2}), \tag{145}$$

where $M' = M - X_i$ and $M'' = M - X_i - X_j$. Using these relations, the law of large numbers and the relation $h^* = h^*_\infty + O(\Delta^{-1})$ where $h^*_\infty = \tilde{\theta}\tanh h^*_\infty$ it is now possible to calculate the limit

$$\lim_{\Delta\to\infty} S(\tilde{\theta}/(\Delta-1)) = \frac{-1 + h^*_\infty \tanh h^*_\infty}{2\cosh^4 h^*_\infty}. \tag{146}$$

This finishes the proof of the second part of the lemma since h_∞ can now be determined by $h_\infty \tanh h_\infty = 1$ and $\tilde{\theta} = h_\infty^2$.

We now show how to deduce the above expression. Let us introduce the following notation,

$$E_0 = \mathbb{E}_+^0\left(\frac{1}{\cosh(\tilde{\theta}M/(\Delta-1))}\right), E_1 = \mathbb{E}_+^0\left(\frac{1}{\cosh(\tilde{\theta}M'/(\Delta-1))}\right), \tag{147}$$

$$E_2 = \mathbb{E}_+^0\left(\frac{1}{\cosh(\tilde{\theta}M''/(\Delta-1))}\right), F_1 = \mathbb{E}_+^0\left(\frac{\tanh(\tilde{\theta}M'/(\Delta-1))}{\cosh(\tilde{\theta}M'/(\Delta-1))}\right), \tag{148}$$

$$F_2 = \mathbb{E}_+^0\left(\frac{\tanh(\tilde{\theta}M''/(\Delta-1))}{\cosh(\tilde{\theta}M''/(\Delta-1))}\right), D = \mathbb{E}_+^0\left(\cosh(\tilde{\theta}M/(\Delta-1))\right). \tag{149}$$

With this in mind and recalling that $\theta = \tilde{\theta}/(\Delta - 1)$ we can write,

$$S(\theta) = \tag{150}$$

$$\frac{\tanh h^*}{2D}\left(\Delta E_1 \tanh h^* - \frac{\tilde{\theta}\Delta}{\Delta - 1}F_1\right) \tag{151}$$

$$-\frac{1}{2\Delta D}\left(1 - \frac{\tilde{\theta}}{\Delta - 1}\text{sech}^2 h^*\right)\left(\Delta E_0 + \Delta(\Delta - 1)\left(E_2 \tanh^2 h^* - \frac{2\tilde{\theta}}{\Delta - 1}F_2 \tanh h^*\right)\right) \tag{152}$$

$$+ O(\Delta^{-1}) \tag{153}$$

$$= \frac{\tanh^2 h^*}{2D}(\Delta E_1 - (\Delta - 1)E_2) - \frac{E_0}{2D} - \frac{\Delta}{\Delta - 1}\frac{F_1}{2D}\tilde{\theta}\tanh h^* + \frac{F_2}{D}\tilde{\theta}\tanh h^* \tag{154}$$

$$+ \frac{E_2}{2D}\tilde{\theta}\text{sech}^2 h^* \tanh^2 h^* + \frac{1}{\Delta - 1}\frac{E_0}{2D}\tilde{\theta}\text{sech}^2 h^* - \frac{1}{\Delta - 1}\frac{F_2}{2D}\tilde{\theta}^2 \tanh h^* \text{sech}^2 h^* + O(\Delta^{-1}). \tag{155}$$

Now notice that expanding the cosh(.) inside E_1 in expression $\Delta E_1 - (\Delta - 1)E_2$ around $M''\tilde{\theta}/(\Delta - 1)$ we can rewrite the same expression as,

$$E_2 - \frac{\Delta}{\Delta - 1}\tilde{\theta}\tanh h^* F_2 + O(\Delta^{-1}). \tag{156}$$

Inserting this in the above expression finally gives us,

$$S(\theta) = \frac{E_2}{2D}\tanh^2 h^* - \frac{\Delta}{\Delta - 1}\frac{F_2}{2D}\tilde{\theta}\tanh^3 h^* - \frac{E_0}{2D} - \frac{\Delta}{\Delta - 1}\frac{F_1}{2D}\tilde{\theta}\tanh h^* + \frac{F_2}{D}\tilde{\theta}\tanh h^* \tag{157}$$

$$+ \frac{E_2}{2D}\tilde{\theta}\text{sech}^2 h^* \tanh^2 h^* + O(\Delta^{-1}). \tag{158}$$

By the law of large numbers we have,

$$\lim_{\Delta\to\infty} M/(\Delta - 1) = \lim_{\Delta\to\infty} M'/(\Delta - 1) = \lim_{\Delta\to\infty} M''/(\Delta - 1) \tag{159}$$

$$= \lim_{\Delta\to\infty} \mathbb{E}_+(X_i)\Big|_{\theta=\tilde{\theta}/(\Delta-1)} = \tanh h^*_\infty, \tag{160}$$

and since all the variables inside the expectations are uniformly bounded, we can take the limit inside all the expectations of our expression for $S(\theta)$. Doing so we get,

$$\lim_{\theta\to\infty} S(\tilde{\theta}/(\Delta - 1)) = \frac{\tanh^2 h^*_\infty}{2\cosh^2 h^*_\infty} - \frac{\tilde{\theta}\tanh^4 h^*_\infty}{2\cosh^2 h^*_\infty} - \frac{1}{2\cosh^2 h^*_\infty} - \frac{\tilde{\theta}\tanh^2 h^*_\infty}{2\cosh^2 h^*_\infty} \tag{161}$$

$$+ \frac{\tilde{\theta}\tanh^2 h^*_\infty}{\cosh^2 h^*_\infty} + \frac{\tilde{\theta}\tanh^2 h^*_\infty}{2\cosh^4 h^*_\infty}. \tag{162}$$

If we now use the relation $h_\infty = \tilde{\theta}\tanh h_\infty$ this expression can be simplified to,

$$\frac{1}{2\cosh^4 h^*_\infty}(-1 + h^*_\infty \tanh h^*_\infty). \tag{163}$$

Finally, we show that there exists a constant $C_{\min}$ such that

$$\lim_{p\to\infty} \sigma_{\min}(Q^*_{SS}) = \sigma_{\min}\left(\lim_{p\to\infty} Q^*_{SS}\right) > C_{\min}.^{9} \tag{164}$$

[9]The equality is guaranteed since the sequence of matrices $\{Q^*_{SS}\}_{p=1}^{\infty}$ have fix finite dimensions.

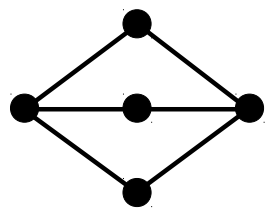


Figure 5: Example of small graph for which the incoherence fails.

First notice that the eigenvalues of $\lim_{p\to\infty} Q^*_{SS}$ are $a-b$ and $a+(\Delta-1)b = c_1 - c_2$. It is immediate to see that $a-b>0$. In addition, since $\Delta(c_1-c_2) = \mathbb{E}_+(M^2/\cosh^2(\theta M)) > 0$ it follows that $c_1 - c_2 > 0$. Hence we can choose $C_{\min} = \min\{a-b, c_1-c_2\} > 0$.

□

D Analysis of $\mathsf{RIr}(\lambda)$ for other families of graphs

As already discussed, the success of RIr is closely related to the incoherence condition. For small graphs, brute force computations allow to explicitly evaluate this condition. For example, consider the reconstruction of the neighborhood of the leftmost node in the graph of Figure 5. The corresponding incoherence parameter takes the for,

$$\|Q^*_{S^CS}{Q^*_{SS}}^{-1}\|_\infty = \frac{3x(1+x^2)}{1+3x^2}, \tag{165}$$

where $x = \tanh\theta$. For $x > x_* \equiv \frac{1}{3}\left(1 - \sqrt[3]{2} + 2^{2/3}\right) \approx 0.44249$ (i.e. for $\theta > \operatorname{atanh}(x_*) \approx 0.475327$) the right hand side is larger than 1, whence the incoherence condition is violated $\|Q^*_{S^CS}{Q^*_{SS}}^{-1}\|_\infty > 1$.

This simple calculation strongly suggests that $\mathsf{RIr}(\lambda)$ fails on the graph of Figure 5 for $\theta > \operatorname{atanh}(x_*)$, although it does not provide a complete proof of this failure. In this appendix we study three classes of graphs of increasing size. We show that with high probability RIr succeeds in reconstructing trees. On the other hand, we show that it fails –for θ large enough– at reconstructing large two-dimensional grids, and that in fails in reconstructing graphs G_p from the toy example in Section 1.1.

D.1 Trees

Lemma D.1. *If G is a tree rooted at r with depth > 1 and node r has degree $\Delta > 1$ then, for this node*

$$\|Q^*_{S^CS}{Q^*_{SS}}^{-1}\|_\infty = \tanh\theta < 1, \tag{166}$$

$\sigma_{\min}(Q^*_{SS}) \geq (1-\tanh^2\theta)/\cosh^2(\theta\Delta)$ *and* $\sigma_{\max}(\mathbb{E}_{G,\theta}(X_S^T X_S)) = 1 + (\Delta-1)\tanh^2\theta$.

Proof. In what follows $\mathbb{E}$ will denote $\mathbb{E}_{G,\theta}$. Consider a node $r' \in S^C$ and let $k \in S$ be the unique node in S that belongs to the shortest path connecting r' to r. Let t be the distance between r' and k. For every $i \in S$ one can write,

$$Q^*_{r'i} = \mathbb{E}(X_{r'}X_i/\cosh^2(\theta M)) = \mathbb{E}(X_{r'}X_k)\mathbb{E}(X_kX_i/\cosh^2(\theta M)) = (\tanh\theta)^t\,\mathbb{E}(X_kX_i/\cosh^2(\theta M)).$$

This equation is still valid if $k = i$. We can thus write that $Q^*_{r'S} = (\tanh\theta)^t\ Q^*_{kS}$ and hence $Q^*_{r'S}(Q^*_{SS})^{-1} = (\tanh\theta)^t e_k$ where e_k is a row vector with all entries equal to zero except k^{th} entry

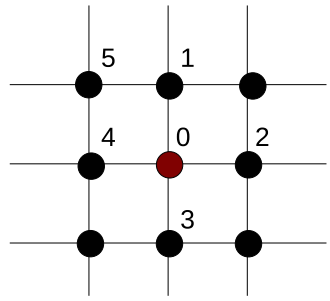


Figure 6: Labeling of the nodes in the grid.

that equals 1. Therefore we can write $\|Q^*_{r'S}(Q^*_{SS})^{-1}\|_1 = (\tanh\theta)^t$. Since there must exist at least one node $r' \in S$ for which the corresponding node k is at distance 1 from S, that is for which $t = 1$, we conclude that $\|Q^*_{S^CS}{Q^*_{SS}}^{-1}\|_\infty = \tanh\theta < 1$.

To prove the spectral bounds first notice that the positive-semidefinite matrix Q^*_{SS} has entries $Q^*_{ij} = (a-b)\delta_{ij} + b$ where $a = \mathbb{E}(1/\cosh^2(\theta M))$ and $b = \mathbb{E}(X_1X_2/\cosh^2(\theta M))$ and where 1 and 2 are any two distinct nodes in S. A matrix of this form has eigenvalues $a-b$ and $a+(\Delta-1)b$. It is not hard to see that $b \geq 0$ and hence

$$\sigma_{\min}(Q^*_{SS}) = a - b = \mathbb{E}((1-X_1X_2)/\cosh^2(\theta M)) \geq \mathbb{E}(1-X_1X_2)/\cosh^2(\theta\Delta). \tag{167}$$

Since $\mathbb{E}(1-X_1X_2) = 1-\tanh^2\theta$ the lower bound follows.

The computation of the value of the maximum eigenvalue value of $\mathbb{E}_{G,\theta}(X_S^TX_S)$ is trivial since this matrix is also of the form $(a-b)\delta_{ij}+b$ with $a=1$ and $b=\tanh^2\theta$.

□

D.2 Two-dimensional grids

Lemma D.2. *If G is a two dimensional grid with periodic boundary conditions (each node connects to its four closest neighbors) then for p large enough $\theta > \theta_c$ we have $\|Q^*_{S^CS}{Q^*_{SS}}^{-1}\|_\infty > 1+\epsilon$ and $\sigma_{\min}(Q^*_{SS}) > C_{\min}$ where θ_c, $\epsilon > 0$ and $C_{\min}$ are independent of p.*

Proof. We shall compute a lower bound on $\|Q^*_{S^CS}{Q^*_{SS}}^{-1}\|_\infty$ by means of a low temperature expansion, i.e. a Taylor expansion in powers of $e^{-\theta}$. We will show that for this lower bound the lemma holds.

Label the central node as node 0, the neighboring nodes as 1, 2, 3 and 4. Denote as node 5 be the common neighbor of node 1 and node 4. Throughout this proof we will denote $\mathbb{E}_{G,\theta}$ by $\mathbb{E}$ and $\mathbb{P}_{G,\theta}$ by $\mathbb{P}$.

First notice that due to the periodic boundary condition there is symmetry along the vertical and horizontal axis in the lattice. Knowing this, matrix Q^*_{SS} can be written as

$$\begin{bmatrix} a & b & c & b \\ b & a & b & c \\ c & b & a & b \\ b & c & b & a \end{bmatrix}, \tag{168}$$

where $a = \mathbb{E}(1/\cosh^2(\theta M))$, $b = \mathbb{E}(X_1X_2/\cosh^2(\theta M))$ and $c = \mathbb{E}(X_1X_3/\cosh^2(\theta M))$, where $M = \sum_{i\in\partial i} X_i$, that is, M is the sum of the variables in the neighborhood of i (i not included). Since we

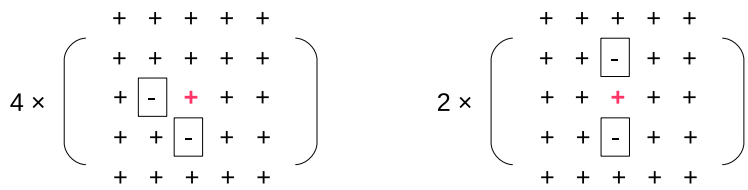


Figure 7: Basic type of configurations for the calculation of $\mathbb{P}(|M|=0)$. The number in front of each picture represents the number of equivalent symmetric configurations that need to be taken into account.

only want to prove a lower bound on $\|Q^*_{S^CS}{Q^*_{SS}}^{-1}\|_\infty$ we only consider the row of $Q^*_{S^CS}$ associated with node 5. This row has the form,

$$\left[\begin{array}{cccc} d & e & e & d \end{array}\right], \tag{169}$$

where $d = \mathbb{E}(X_1X_5/\cosh^2(\theta M))$ and $e = \mathbb{E}(X_2X_5/\cosh^2(\theta M))$. To compute the low temperature expansions of each of these quantities we first write,

$$\begin{aligned} a &= \mathbb{P}(|M|=0) + \frac{1}{\cosh^2 2\theta}\mathbb{P}(|M|=2) + \frac{1}{\cosh^2 4\theta}\mathbb{P}(|M|=4), && (170)\\ \mathbb{E}\left(\frac{X_iX_j}{\cosh^2\theta M}\right) &= [\mathbb{P}(|M|=0, X_iX_j=1) - \mathbb{P}(|M|=0, X_iX_j=-1)] && (171)\\ &+ \frac{1}{\cosh^2 2\theta}[\mathbb{P}(|M|=2, X_iX_j=1) - \mathbb{P}(|M|=2, X_iX_j=-1)] && (172)\\ &+ \frac{1}{\cosh^2 4\theta}[\mathbb{P}(|M|=4, X_iX_j=1) - \mathbb{P}(|M|=4, X_iX_j=-1)]. && (173) \end{aligned}$$

The problem thus resumes to the computation of the above probabilities. We will exemplify the calculation of the low temperature expansion of $\mathbb{P}(|M|=0)$, the calculation of the expansion for the other terms follows in a similar fashion.

Let $\mathcal{H}(\underline{x}) = \sum_{(ij)\in E} x_ix_j$, $\mathcal{H}_{\max} = \max_{\underline{x}} \mathcal{H}(\underline{x}) = |E|$ and $\delta\mathcal{H}(\underline{x}) = \mathcal{H}(\underline{x}) - \mathcal{H}_{\max} = -2\mathcal{P}(\underline{x})$ where $\mathcal{P}(\underline{x})$ is the length of the boundary separating positive spins from negative spins in configuration $\underline{x}$. Then,

$$\begin{aligned} \mathbb{P}(|M|=0) &= \frac{2}{Z}\sum_{\{\underline{x}:x_0=1, M=0\}} e^{\theta\mathcal{H}(\underline{x})} && (174)\\ &= \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}\sum_{s\geq 4}\sum_{\{\underline{x}:x_0=1, M=0, \mathcal{P}=s\}} e^{-2\theta s}. && (175) \end{aligned}$$

The term $2e^{\theta\mathcal{H}_{\max}}/Z$ appears in all a, b, c, d and e and thus is irrelevant for the computation of $[Q^*_{S^CS}{Q^*_{SS}}^{-1}]_5$. Since only configurations with zero magnetization contribute to the sum there are two basic types of configurations we need to consider, both of which must have exactly two neighbors of node 0 with negative spin. These are represented in figure 7. Starting from these two basic states we need to consider the first few lowest energy configurations. To help the counting there are two parameters that we keep track of: the number of negative spins, t, and the perimeter of the boundary, s. The first type of state produces the counting expressed in table 1. The associated configurations are represented in figure 8.

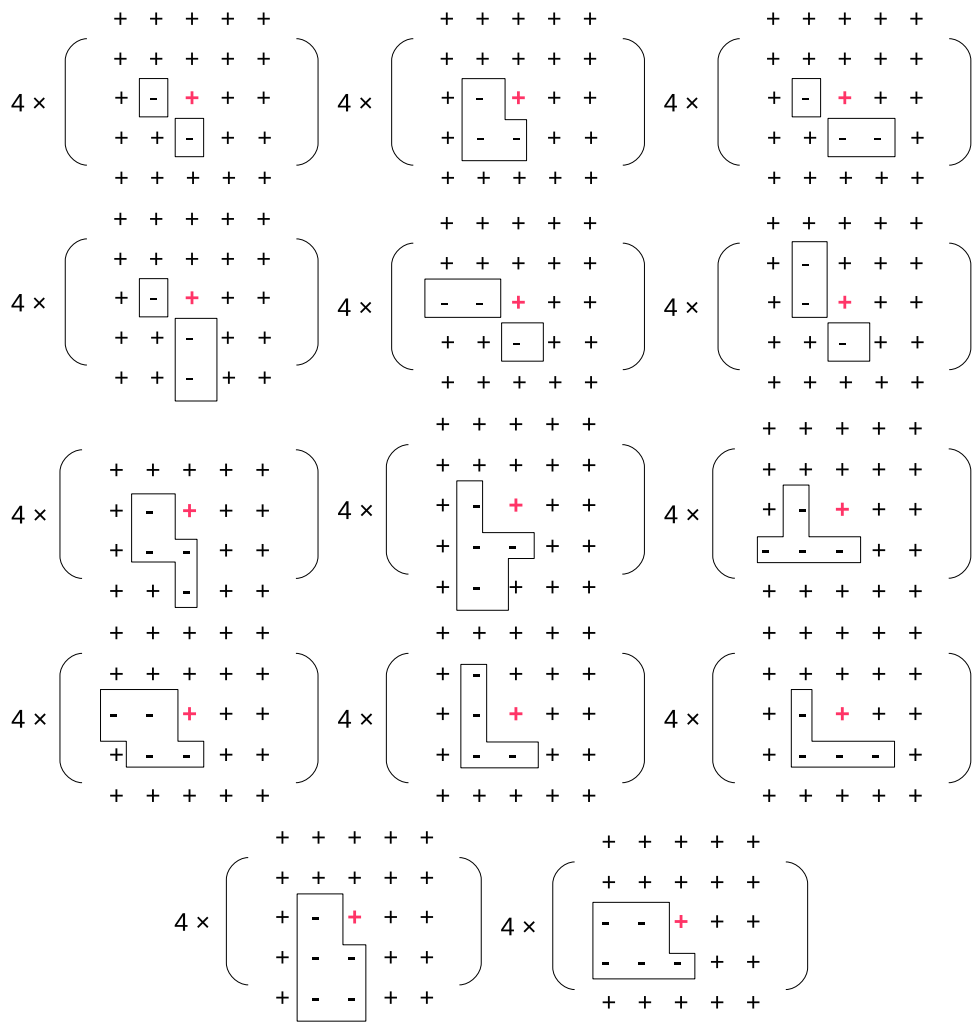


Figure 8: Configurations derived from first basic type of configuration for the calculation of $\mathbb{P}(|M| = 0)$.

Table 1: Low energy states from first basic configuration for low temperature expansion of $\mathbb{P}(|M| = 0)$

Negative spins, t	Boundary perimeter, s	Number of states
2	8	4×1
3	8	4×1
3	10	4×4
4	10	4×6
5	10	4×2

For the second type of basic configuration the counting is in table 2 and the associated configurations in figure 9.

Table 2: Low energy states from second basic configuration for low temperature expansion of $\mathbb{P}(|M| = 0)$

Negative spins, t	Boundary perimeter, s	Number of states
2	8	4×1
3	10	4×6

We can thus write,

$$\mathbb{P}(|M| = 0) = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (10e^{-16\theta} + 60e^{-20\theta} + O(e^{-24\theta})). \tag{176}$$

For the expansion of $\mathbb{P}(|M| = 2)$ we also have two basic states types from which all the other ones are built. The first type has only one negative spin in the neighborhood of node 0 and the second type has 3 negative spins in the neighborhood of node 0. See figure 10.

The counting of states derived from the first basic state type and second basic state type are recorded in tables 3 and 4 respectively.

Table 3: Low energy states from first basic configuration for calculation of $\mathbb{P}(|M| = 2)$

Negative spins, t	Boundary perimeter, s	Number of states
1	4	4×1
2	6	4×3
2	8	$4 \times$ ($\|E\|$ - 8)
3	8	4×10
4	8	4×2

We can thus write,

$$\mathbb{P}(|M| = 2) = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (4e^{-8\theta} + 12e^{-12\theta} + O(e^{-16\theta})). \tag{177}$$

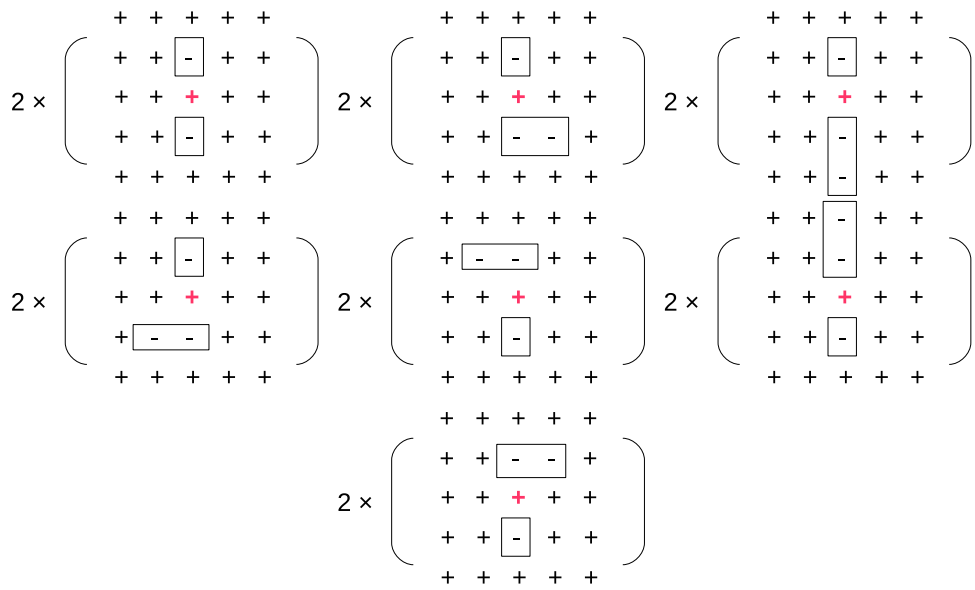


Figure 9: Configurations derived from second basic type of configuration for the calculation of $\mathbb{P}(|M| = 0)$.

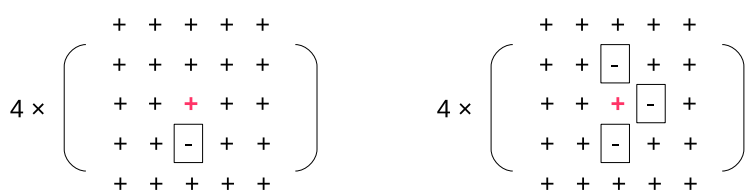


Figure 10: Basic type of configurations for the calculation of $\mathbb{P}(|M| = 2)$. The number in front of each picture represents the number of equivalent symmetric configurations that need to be taken into account.

Table 4: Low energy states from second basic configuration for calculation of $\mathbb{P}(|M| = 2)$

Negative spins, t	Boundary perimeter, s	Number of states
3	12	4×1

For the expansion of $\mathbb{P}(|M| = 4)$ we again have two basic states types from which all the other ones are built. The first type has all spins positive in the neighborhood of node 0 and the second type has all spins negative in the neighborhood of node 0. The counting of states in printed in table 5.

Table 5: Low energy states for calculation of $\mathbb{P}(|M| = 4)$

Negative spins, t	Boundary perimeter, s	Number of states
0	0	1
1	4	$\|E\|$ - 5
4	16	1

We thus have,

$$\mathbb{P}(|M| = 4) = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (4e^{-8\theta} + 12e^{-12\theta} + O(e^{-16\theta})). \tag{178}$$

Using the expansion $1/\cosh^2(x) = 4e^{-2x}(1 - 2e^{-2x} + 3e^{-4x} + O(e^{-8x}))$ we can finally write,

$$a = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (4e^{-8\theta} + 16e^{-12\theta} + O(e^{-16\theta})). \tag{179}$$

For the probabilities involved in the calculation of b we get the following expansions,

$$\mathbb{P}(|M| = 0, X_1X_2 = 1) = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (4e^{-16\theta} + 24e^{-20\theta} + O(e^{-24\theta})), \tag{180}$$

$$\mathbb{P}(|M| = 0, X_1X_2 = -1) = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (6e^{-16\theta} + 38e^{-20\theta} + O(e^{-24\theta})), \tag{181}$$

$$\mathbb{P}(|M| = 2, X_1X_2 = 1) = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (2e^{-8\theta} + 6e^{-12\theta} + O(e^{-16\theta})), \tag{182}$$

$$\mathbb{P}(|M| = 2, X_1X_2 = -1) = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (2e^{-8\theta} + 6e^{-12\theta} + O(e^{-16\theta})), \tag{183}$$

$$\mathbb{P}(|M| = 4, X_1X_2 = 1) = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (1 + (|E| - 5)e^{-8\theta} + O(e^{-12\theta})), \tag{184}$$

$$\mathbb{P}(|M| = 4, X_1X_2 = -1) = 0, \tag{185}$$

and putting everything together we obtain,

$$b = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (4e^{-8\theta} + (4|E| - 30)e^{-16\theta} + O(e^{-20\theta})). \tag{186}$$

For the probabilities involved in the calculation of c we get the following expansions,

$$\mathbb{P}(|M|=0, X_1X_3=1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(2e^{-16\theta}+12e^{-20\theta}+O(e^{-24\theta})), \tag{187}$$

$$\mathbb{P}(|M|=0, X_1X_3=-1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(8e^{-16\theta}+48e^{-20\theta}+O(e^{-24\theta})), \tag{188}$$

$$\mathbb{P}(|M|=2, X_1X_3=1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(2e^{-8\theta}+6e^{-12\theta}+O(e^{-16\theta})), \tag{189}$$

$$\mathbb{P}(|M|=2, X_1X_3=-1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(2e^{-8\theta}+6e^{-12\theta}+O(e^{-16\theta})), \tag{190}$$

$$\mathbb{P}(|M|=4, X_1X_3=1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(1+(|E|-5)e^{-8\theta}+O(e^{-12\theta})), \tag{191}$$

$$\mathbb{P}(|M|=4, X_1X_3=-1) = 0, \tag{192}$$

and putting everything together we obtain,

$$c = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(4e^{-8\theta}+(4|E|-34)e^{-16\theta}+O(e^{-20\theta})). \tag{193}$$

For the probabilities involved in the calculation of d we get the following expansions,

$$\mathbb{P}(|M|=0, X_1X_5=1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(6e^{-16\theta}+38e^{-20\theta}+O(e^{-24\theta})), \tag{194}$$

$$\mathbb{P}(|M|=0, X_1X_5=-1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(4e^{-16\theta}+19e^{-20\theta}+O(e^{-24\theta})), \tag{195}$$

$$\mathbb{P}(|M|=2, X_1X_5=1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(3e^{-8\theta}+9e^{-12\theta}+O(e^{-16\theta})), \tag{196}$$

$$\mathbb{P}(|M|=2, X_1X_5=-1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(e^{-8\theta}+3e^{-12\theta}+O(e^{-16\theta})), \tag{197}$$

$$\mathbb{P}(|M|=4, X_1X_5=1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(1+(|E|-6)e^{-8\theta}+O(e^{-12\theta})), \tag{198}$$

$$\mathbb{P}(|M|=4, X_1X_5=-1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(e^{-8\theta}+O(e^{-12\theta})), \tag{199}$$

and putting everything together we obtain,

$$d = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(4e^{-8\theta}+8e^{-12\theta}+(4|E|-30)e^{-16\theta}+O(e^{-20\theta})). \tag{200}$$

For the probabilities involved in the calculation of e we get the following expansions,

$$\mathbb{P}(|M|=0, X_2X_5=1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(4e^{-16\theta}+22e^{-20\theta}+O(e^{-24\theta})), \tag{201}$$

$$\mathbb{P}(|M|=0, X_2X_5=-1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(6e^{-16\theta}+38e^{-20\theta}+O(e^{-24\theta})), \tag{202}$$

$$\mathbb{P}(|M|=2, X_2X_5=1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(3e^{-8\theta}+7e^{-12\theta}+O(e^{-16\theta})), \tag{203}$$

$$\mathbb{P}(|M|=2, X_2X_5=-1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(e^{-8\theta}+5e^{-12\theta}+O(e^{-16\theta})), \tag{204}$$

$$\mathbb{P}(|M|=4, X_2X_5=1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(1+(|E|-6)e^{-8\theta}+O(e^{-12\theta})), \tag{205}$$

$$\mathbb{P}(|M|=4, X_2X_5=-1) = \frac{2}{Z}e^{\theta\mathcal{H}_{\max}}(e^{-8\theta}+O(e^{-12\theta})), \tag{206}$$

and putting everything together we obtain,

$$e = \frac{2}{Z} e^{\theta \mathcal{H}_{\max}} (4e^{-8\theta} + 8e^{-12\theta} + (4|E| - 46)e^{-16\theta} + O(e^{-20\theta})). \tag{207}$$

Using the expansions for a, b, c, d and e and computing the series expansion of $[Q^*_{S^C S} {Q^*_{SS}}^{-1}]_5$ in powers of $e^{-\theta}$ we finally obtain,

$$\|[Q^*_{S^C S} {Q^*_{SS}}^{-1}]\|_\infty \geq \|[Q^*_{S^C S} {Q^*_{SS}}^{-1}]_5\|_1 = 1 + e^{-4\theta} + O(e^{-8\theta}). \tag{208}$$

Following the ideas of [31] one can then show that the above formal expansion converges (a priori it could be case that one of the higher order terms would depend on $|E|$). This finishes the first part of the proof.

We now prove that there exists $C_{\min} > 0$ such that $\lim_{p\to\infty} \sigma(Q^*_{SS}) > C_{\min}$. This will prove the second part of the theorem. First notice that the eigenvalues of Q^*_{SS} are $\{a-c, a+2b+c, a-2b+c\}$. Now notice that,

$$a - c = \mathbb{E}\left(\frac{1 - X_1 X_2}{\cosh^2(\theta M)}\right), \tag{209}$$

$$a + 2b + c = \frac{1}{4}\mathbb{E}\left(\frac{M^2}{\cosh^2(\theta M)}\right), \tag{210}$$

$$a - 2b + c = \frac{1}{4}\mathbb{E}\left(\frac{(X_1 + X_3 - X_2 - X_4)^2}{\cosh^2(\theta M)}\right). \tag{211}$$

where for $a + 2b + c$ and $a - 2b + c$ we made use of the symmetry of the lattice. Since $1 - X_1 X_2$, M and $X_1 + X_3 - X_2 - X_4$ only depend on a fixed finite number of spins, and since $\theta < \infty$, there is a positive probability, independent of p, of their being non-zero. Hence, all eigenvalues of Q^*_{SS} are strictly positive even as $p \to \infty$. □

D.3 Graphs G_p from the toy example

In this section we show that $\mathsf{Rlr}(\lambda)$ fails to reconstruct the graphs G_p defined in Section 1.1 (see Figure 1) for all λ when θ is large enough. Note that this differs from previous analysis in the sense that we do not require that $\lambda \to 0$. We also show that this 'critical' θ behaves like Δ^{-1} for large Δ. Our analysis is based on numerical evaluation of functions for which explicit analytic expressions can be given along the lines of Section A. Hence, our argument should be understood as a sketch of a proof.

The success of $\mathsf{Rlr}(\lambda)$ is dictated by the behavior of $L(\underline{\theta}_{r,\cdot}; \{x^{(\ell)}\}_{\ell=1}^n)$ when n is large. In fact, it is easy to use concentration inequalities to show that the solution of Rlr for finite n converges with high probability to the minima of $L_\infty(\underline{\theta}) + \lambda\|\underline{\theta}\|_1$ where $L_\infty(\underline{\theta}) \equiv \lim_{n\to\infty} L(\underline{\theta}_{r,\cdot}; \{x^{(\ell)}\}_{\ell=1}^n)$.

If $\lambda \to 0$ as $n \to \infty$, we have seen that the success of Rlr is dictated by the incoherence condition, which in turn is determined by the Hessian of $L_\infty(\underline{\theta})$. It is not hard to see that for this family of graphs, $\|Q^*_{S^C S} {Q^*_{SS}}^{-1}\|_\infty$ is increasing with p. For $p = 5$, Eq. (165) tells us that the incoherence condition will be violated for θ high enough. Hence, by Lemma 4.1, Rlr will fail for all G_p ($p \geq 5$) when $\lambda \to 0$ as $n \to \infty$. The question now is: how does $\mathsf{Rlr}(\lambda)$ behave if $\lambda \to 0$ does not hold?

If $\lambda >$ constant > 0, the success of Rlr is dictated by the minima of $L_\infty(\underline{\theta}) + \lambda\|\underline{\theta}\|_1$. For this specific family of graphs, it is also not hard to see that for $0 < \theta < \infty$, L_∞ is strictly convex and that due to symmetry the unique minimum of $L_\infty(\underline{\theta}) + \lambda\|\underline{\theta}\|_1$ must satisfy $\hat{\theta}_{13} = \hat{\theta}_{14} = \cdots = \hat{\theta}_{1p}$ for any λ. This allows us to consider $L_\infty(\underline{\theta})$ as a function of only two parameters. We call it

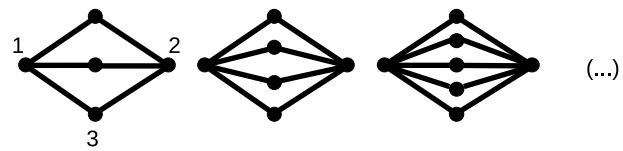


Figure 11: For this family of graphs of increasing maximum degree Δ $\mathsf{Rlr}(\lambda)$ will fail for any $\lambda > 0$ if $\theta > K/\Delta$, where K is a large enough constant.

$L'(\theta_{13}, \theta_{12}) \equiv L(\theta_{12}, \theta_{13}, \theta_{13}, ..., \theta_{13})$. Now, the problem of understanding Rlr for $\lambda > 0$, large n and any p becomes tractable and associated to understanding the following problem,

$$\min_{\theta_{13},\theta_{12}} L'(\theta_{13}, \theta_{12}) + \lambda(p-2)|\theta_{13}| + \lambda|\theta_{12}|. \tag{212}$$

We can analyze this optimization problem by solving it numerically. Figure 12 shows the solution path of this problem as a function of λ for $p = 5$ and for different values of θ.

From the plots we see that for high values of θ, Rlr will never yield a correct reconstruction (unless we assume $\lambda = 0$) since for these θs all curves are strictly above the horizontal axis, that is, $\hat{\theta}_{12} > 0$. However, if θ is bellow a certain value, call it θ_T ($\theta_T \approx 0.61$ for graph G_5), then there are solution that yield a correct reconstruction if we choose values of $\lambda > 0$. In fact, for $\theta < \theta_T$ all curves exhibit a portion (above a certain λ) that have $\hat{\theta}_{12} = 0$ and $\hat{\theta}_{13} > 0$. That is, for $\theta < \theta_T$, Rlr makes a correct structural reconstruction. If we make θ even smaller then the curves identify themselves with the horizontal axis. We call by θ_L the value of θ below which this occurs.

We again note that all previous considerations were made in the limit when $n \to \infty$. For high finite n, with high probability the solution curves will not be the ones plotted but rather be random fluctuations around these. For $\lambda = 0$, finite n and $\theta > \theta_L$, the solution curves will no longer start from $\underline{\hat{\theta}} = \underline{\theta}^* = (\theta, 0)$ but will have a positive non vanishing probability of having $\hat{\theta}_{12} > 0$. This reflects the fact that for finite n the success of $\mathsf{Rlr}(\lambda)$ requires λ to be positive. However, for $\theta < \theta_L$ and $\lambda > 0$ such that we are in the region where the curves for $n = \infty$ are identically zero, the curves for finite n will have an increasing probability of being identically zero too. Thus, for these values of λ and θ, the probability of successful reconstruction will tend to 1 as $n \to \infty$. From the plots we also conclude that, unless the whole curve (for $n = \infty$) is identified with zero, $\mathsf{Rlr}(\lambda)$ restricted to the assumption $\lambda \to 0$ will fail with positive non vanishing probability for finite n. For $\theta < \theta_L$, when the curves (for $n = \infty$) become identically zero, there will be a scaling of λ with n to zero that will allow for a probability of success converging to 1 as $n \to \infty$.

Requiring $\lambda \to 0$ makes θ_L be the critical value above which reconstruction with Rlr fails. This is the scenario in which we studied Rlr in section 2.2.2. In fact, θ_L coincides with the value above which $\|Q^*_{S^cS}{Q^*_{SS}}^{-1}\|_\infty > 1$. For this family of graphs we thus conclude that the true condition required for successful reconstruction is not $\|Q^*_{S^cS}{Q^*_{SS}}^{-1}\|_\infty < 1$ but rather that $\theta < \theta_T$. Surprisingly, for graphs in G_p this condition coincides with $\mathbb{E}_{G,\theta}(X_1X_3) > \mathbb{E}_{G,\theta}(X_1X_2)$, i.e. the correlation between neighboring nodes must be bigger than that between non-neighboring nodes. Notice that this condition is in fact the condition required for Thr to work. Consequently, for this family of graphs, the thresholding algorithm will always have a working range in terms of θ larger than that of Rlr, when restricted to $\lambda \to \infty$. In fact, a simple calculation using the local weak convergence used in proving Lemma 4.3 shows that with high probability, for large random regular graphs, the correlation between neighboring nodes is always strictly greater than between non-neighboring nodes.

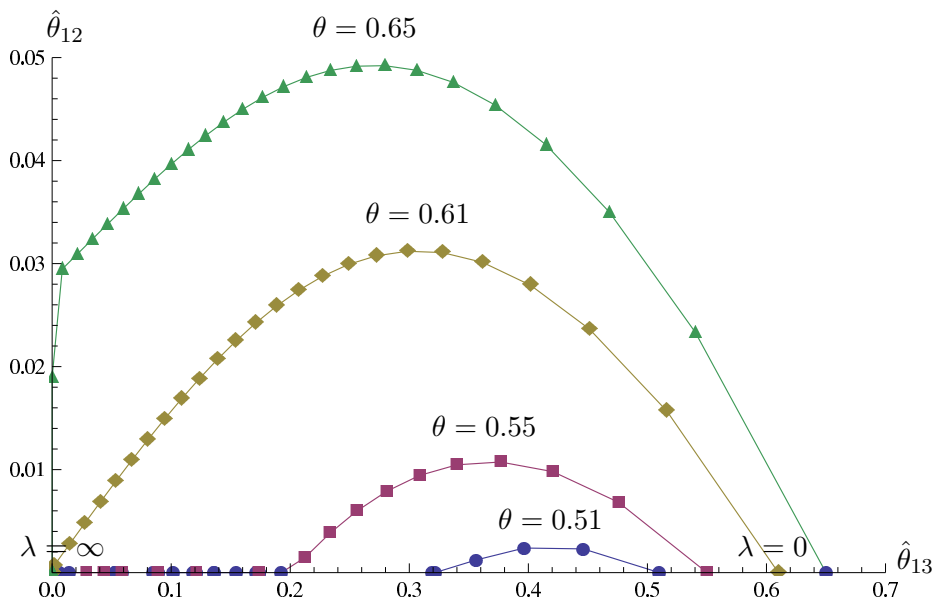


Figure 12: Solution curves of $\mathsf{Rlr}(\lambda)$ as a function of λ for different values of θ and $p = 5$. Along each curve, λ increases from right to left. Plot points separated by $\delta\lambda = 0.05$ are included to show the speed of the parameterization with λ. For $\lambda \to \infty$ all curves tend to the point $(0, 0)$. Remark: Curves like the one for $\theta = 0.55$ are identically zero above a certain value of λ.

This shows that the thresholding algorithm has as operation range $\theta \in (0, \infty)$ for random regular graphs, compared to $\theta \in (0, \theta_L)$ for Rlr.

We will now prove that for large enough $\Delta = p - 2$ there is a unique $\theta_T(\Delta)$ (solution of $\mathbb{E}_{G,\theta,\Delta}(X_1X_3) = \mathbb{E}_{G,\theta,\Delta}(X_1X_2)$) that scales like $1/\Delta$ and above which $\mathbb{E}_{G,\theta,\Delta}(X_1X_3) < \mathbb{E}_{G,\theta,\Delta}(X_1X_2)$. Let 1 and 2 be the two nodes with degree greater than 2 and let 3 be any other node (of degree 2), see Figure 11. Define $x_\Delta = \mathbb{E}_{G,\theta,\Delta}(X_1X_2)$ and $y_\Delta = \mathbb{E}_{G,\theta,\Delta}(X_1X_3)$. It is not hard to see that,

$$x_{\Delta+1} = \frac{x_\Delta + \tanh^2\theta}{1 + \tanh^2\theta\, x_\Delta} \qquad y_{\Delta+1} = \frac{\tanh\theta\, x_\Delta + \tanh\theta}{1 + \tanh^2\theta\, x_\Delta}. \tag{213}$$

From these expression we see that the condition $x_\Delta(\theta) > y_\Delta(\theta)$ is equivalent to $x_{\Delta-1}(\theta) > \tanh\theta$. Remembering that expectations on the Ising model (1) can be computed from subgraphs of G, [35], an easy calculation shows that,

$$x_\Delta(\theta) = \frac{(1 + z(\theta))^\Delta - (1 - z(\theta))^\Delta}{(1 + z(\theta))^\Delta + (1 - z(\theta))^\Delta}, \tag{214}$$

where $z(\theta) = \tanh^2(\theta)$. Since $x_\Delta \to 1$ with Δ then any θ_T also goes to 0 with Δ and attending to the slope and concavity of $x_\Delta(\theta)$ and $\tanh(\theta)$ for small θ it is easy to see that for large Δ there will exist a unique solution $\theta_T(\Delta)$. Furthermore, the condition $x_{\Delta+1}(\theta) = y_{\Delta+1}(\theta)$ can now be written like,

$$\sqrt{z(\theta)} = \frac{(1 + z(\theta))^\Delta - (1 - z(\theta))^\Delta}{(1 + z(\theta))^\Delta + (1 - z(\theta))^\Delta}. \tag{215}$$

Assuming $z = K\Delta^{-\gamma}$, multiplying both sides of the previous equation by $\Delta^{\gamma/2}$ and taking the limit when $\Delta \to \infty$ we obtain,

$$\sqrt{K} = \lim_{\Delta\to\infty} \Delta^{\gamma/2} \tanh(K\Delta^{1-\gamma}), \tag{216}$$

which will result in a non trivial relation for K only if $\gamma = 2$. In this case we get $K^{1/2} = K$ and thus for any $\epsilon > 0$, if Δ is sufficiently high, there will be a (unique) solution of (215) inside the interval $[(1-\epsilon)/\Delta^2, (1+\epsilon)/\Delta^2]$. Since $z(\theta) = \tanh^2(\theta)$ then $\theta_T(\Delta)$ scales likes $1/\Delta$ as we wanted to prove.

References

[1] K. Huang, *Statistical Mechanics*, Wiley, New York, 1987.

[2] G. Grimmett, *The Random-Cluster Model*, Springer-Verlag, New York, 2009.

[3] J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proceedings of the National Academy of Sciences of the USA, 1982, Vol. 79, 2554–2558.

[4] G. Hinton and T. Sejnowski, *Analyzing Cooperative Computation*, Proc. of the 5th Annual Congress of the Cognitive Science Society, Rochester, NY, 1983.

[5] E. Lehmann and G. Casella, *Theory of Point Estimation*, Springer, New York, 1998.

[6] M. Wainwright, *Information-Theoretic Limits on Sparsity Recovery in the High-Dimensional and Noisy Setting*, Information Theory, IEEE Transactions on Information Theory, Vol. 55, 2009, 5728–5741.

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

[8] P. Netrapalli, S. Banerjee, S. Sanghavi and S. Shakkottai, *Greedy Learning of Markov Network Structure*, Proc. of Allerton Conf. on Communication, Control and Computing, 2010.

[9] 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.

[10] 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, 2008, 343–356.

[11] P. Ravikumar, M. Wainwright and J. Lafferty, *High-Dimensional Ising Model Selection Using l1-Regularized Logistic Regression*, The Annals of Statistics, Vol. 38, 2010, 1287–1319.

[12] H. Georgii, *Gibbs Measures and Phase Transitions*, Walter de Gruyter, 1988.

[13] E. Mossel and A. Sly, *Exact Thresholds for Ising-Gibbs Samplers on General Graphs*, 2009.

[14] A. Gerschenfeld and A. Montanari, *Reconstruction for Models on Random Graphs*, Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, 2007.

[15] M. Jerrum and A. Sinclair, *Polynomial-time approximation algorithms for the Ising model*, SIAM Journal on Computing, Vol. 22, 1993.

[16] A. Sly, *Computational Transition at the Uniqueness Threshold*, 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, 2010,287–296.

[17] D. Ackley, G. Hinton, and T. Sejnowski, *A Learning Algorithm for Boltzmann Machines*, Cognitive Science 9, 1985.

[18] G. Hinton, S. Osindero and Y. Teh, *A fast learning algorithm for deep belief nets*, Neural Computation 18, 2006.

[19] N. Meinshausen and P. Bühlmann,*High dimensional graphs and variable selection with the lasso*, Annals of Statistics, Vol. 3, 2006,

[20] J. Friedman, T. Hastie and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics 9, 2008.

[21] A. Anandkumar, V. Tan, and A. Willsky, *High-Dimensional Structure Learning of Ising Models on Sparse Random Graphs*, arXiv:1011.0129, 2010.

[22] S. Cocco and R. Monasson, *Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data*, Physical Review Letters, Vol. 106, 2011.

[23] I. Csiszar and Z. Talata, *Consistent estimation of the basic neighborhood structure of Markov random fields*, The Annals of Statistics, 2006, Vol. 1, 123–145.

[24] N. Friedman, I. Nachman, and D. Peer, *Learning Bayesian network structure from massive datasets: The sparse candidate algorithm*, 1999, 206–215.

[25] 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.

[26] 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, 2008, Vol. 9, 485–516.

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

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

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

[30] C. Domb and A. J. Guttmann, *Low-temperature series for the Ising model*, J. Phys., 1970.

[31] J. Lebowitz and A. Mazel, *Improved Peierls Argument for High-Dimensional Ising Models*, Journal of Statistical Physics, 1998, Vol. 90, 1051–1059.

[32] A. Dembo and A. Montanari, *Ising Models on Locally Tree Like Graphs*, arXiv:0804.4726v2 [math.PR], 2008.

[33] A. Montanari, E. Mossel and A. Sly, *The weak limit of Ising models on locally tree-like graphs*, Probability Theory and Related Fields, 2010.

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

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

[36] R. Griffiths, *Correlations in Ising ferromagnets*, Journal of Mathematical Physics, 1967, Vol. 8, 478.