On Model Selection Consistency of Lasso for High-Dimensional Ising Models

Xiangming Meng
The University of Tokyo

Tomoyuki Obuchi
Kyoto University

Yoshiyuki Kabashima
The University of Tokyo

Abstract

We theoretically analyze the model selection consistency of least absolute shrinkage and selection operator (Lasso), *both with and without post-thresholding*, for high-dimensional Ising models. For random regular (RR) graphs of size p with regular node degree d and uniform couplings θ_0, it is rigorously proved that Lasso *without post-thresholding* is model selection consistent in the whole paramagnetic phase with the same order of sample complexity $n = \Omega(d^3 \log p)$ as that of ℓ_1-regularized logistic regression (ℓ_1-LogR). This result is consistent with the conjecture in *Meng, Obuchi, and Kabashima 2021* [Meng et al., 2021] using the non-rigorous replica method from statistical physics and thus complements it with a rigorous proof. For general tree-like graphs, it is demonstrated that the same result as RR graphs can be obtained under mild assumptions of the dependency condition and incoherence condition. Moreover, we provide a rigorous proof of the model selection consistency of Lasso *with post-thresholding* for general tree-like graphs in the paramagnetic phase without further assumptions on the dependency and incoherence conditions. Experimental results agree well with our theoretical analysis.

1 Introduction

Ising model [Ising, 1925] is one renowned binary undirected graphical models (also known as Markov random fields (MRFs)) [Wainwright and Jordan, 2008, Koller and Friedman, 2009, Mezard and Montanari, 2009] with wide applications in various scientific disciplines such as social networking [McAuley and Leskovec, 2012], gene network analysis [Marbach et al., 2012, Krishnan et al., 2020], and protein interactions [Morcos et al., 2011, Liebl and Zacharias, 2021], just to name a few. Given an undirected graph $G = (V, E)$, where $V = \{1, ..., p\}$ is a collection of nodes associated with the binary spins $X = (X_i)_{i=1}^p$ and $E = \{(r, t) \,|\theta^*_{rt} \neq 0\}$ is a collection of undirected edges that specify the pairwise interactions $\theta^* = (\theta^*_{rt})_{r \neq t}$, the joint probability distribution of an Ising model has the following form

$$\mathbb{P}_{\theta^*}(x) = \frac{1}{Z(\theta^*)} \exp\Big\{ \sum_{r \neq t} \theta^*_{rt} x_r x_t \Big\}, \tag{1}$$

where $Z(\theta^*) = \sum_x \exp\Big\{\sum_{r \neq t} \theta^*_{rt} x_r x_t\Big\}$ is the partition function. In general, there are also external fields but here they are assumed to be zero for simplicity. Importantly, the conditional independence between $X = (X_i)_{i=1}^p$ can be well captured by the associated graph G [Wainwright and Jordan, 2008, Koller and Friedman, 2009] and hence one fundamental problem, namely Ising model selection (also widely known as the inverse Ising problem. Please refer to [Nguyen et al., 2017] for a nice review), is to recover the underlying graph structure (edge set E) of G from a collection of n i.i.d. samples $\mathfrak{X}_n := \big\{x^{(1)}, \ldots, x^{(n)}\big\}$, where $x^{(i)} \in \{-1, +1\}^p$ represents the i-th sample. To address this fundamental problem, a variety of methods have been proposed over the past several decades in various fields [Tanaka, 1998, Kappen and Rodríguez, 1998, Ricci-Tersenghi, 2012, Wainwright et al., 2007, Höfling and Tibshirani, 2009, Ravikumar et al., 2010, Decelle and Ricci-Tersenghi, 2014, Bresler, 2015, Vuffray et al., 2016, Lokhov et al., 2018]. Notably, under the framework of the pseudo-likelihood (PL) [Besag, 1975], both ℓ_1-regularized logistic regression (ℓ_1-LogR) [Ravikumar et al., 2010] and ℓ_1-regularized interaction screening estimator (RISE) [Vuffray et al., 2016, Lokhov et al., 2018] are the two most popular methods in reconstructing the graph structure and the number of samples required is even near-optimal with respect to (w.r.t.) previously established information-theoretic lower-bound [Santhanam and Wainwright, 2012].

In this paper, we consider the well-known least absolute shrinkage and selection operator (Lasso) [Tibshirani, 1996] for Ising model selection. At first sight, one might even

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

doubt its suitability for this problem since apparently the Ising snapshots are binary data generated in a nonlinear manner while Lasso is (presumably) used for continuous data with linear regression. In fact, the idea of using linear regression for binary data is not as outrageous (or naive) as one might imagine [Brillinger, 1982, Dobriban and Wager, 2018, Erdogdu et al., 2019], and perhaps surprisingly, sometimes linear regression even outperforms logistic regression as demonstrated in [Gomila, 2021]. Indeed, if our goal is to make predictions of new outcomes, say binary classification, then linear regression might not be a good choice since it is easily prone to out-of-bound forecasts[1]. However, when it comes to other goals such as estimating variables or causal effects [Gomila, 2021], the answer becomes highly nontrivial. For Ising model selection, the goal is not about making predictions of new binary outcomes, but rather inferring the graph structure and thus deciphering the underlying conditional independence between different variables. Hence, given the popularity of Lasso, it is of both practical and theoretical significance to study the (mis-specified) Lasso's *model selection consistency* for the nonlinear Ising models, i.e., under what conditions Lasso can (or cannot) successfully recover the true structure of Ising model. While several early studies [Bento and Montanari, 2009, Lokhov et al., 2018, Meng et al., 2020, Meng et al., 2021] have implied Lasso's potential consistency for Ising model selection, a rigorous theoretical analysis has still largely remained unresolved.

1.1 Our Contributions

We theoretically analyze the model selection consistency of Lasso, *both with and without post-thresholding*, for Ising models in the high-dimensional ($n \ll p$) regime, where the number of vertices $p = p(n)$ may also scale as a function of the sample size n. The paramagnetic phase of Ising models is considered where the coupling strength is relatively small so that the expectation of the magnetization $m := \frac{1}{p}\sum_{i=1}^{p} x_i$ is zero [Nishimori, 2001, Mezard and Montanari, 2009]. Our main contributions are summarized as follows.

(a) For random regular (RR) graphs with regular node degree d and uniform active couplings $\theta^*_{r,t} = \theta_0, \forall (r,t) \in E$, in the paramagnetic phase, i.e., $(d-1)\tanh\theta_0 < 1$, we prove that Lasso without post-thresholding is model selection consistent for Ising models, and remarkably, the required sample complexity has the same scaling order as that of ℓ_1-LogR. (Theorem 1)

(b) For general tree-like graphs, under mild assumptions of the *dependency condition* and *incoherence condition*, it is proved that Lasso without post-thresholding is still model selection consistent for Ising models with the same order of sample complexity as that of ℓ_1-LogR. (Theorem 2)

(c) For general tree-like graphs, we not only obtain an upper bound of the reconstructed square error of Lasso, but also prove that, with some proper post-thresholding, Lasso is model selection consistent with the same order of sample complexity as that of ℓ_1-LogR and RISE without any further assumptions on the *dependency* and *incoherence conditions*. (Theorems 3 and 4)

Remark 1: It is worth strengthening that in this paper we focus on Lasso *both with and without post-theresholding*.

Remark 2: Given the wide popularity and efficiency of Lasso, our analysis not only provides a theoretical backing for its practical use, but also deepens our understanding of learning Ising models using Lasso. Previously, it has long been believed that the success of Lasso for Ising model selection (approximately) happens only when $\theta^*_{r,t} \to 0, \forall (r,t) \in E$ so that the square loss of Lasso is similar to the logistic loss of ℓ_1-LogR [Lokhov et al., 2018]. However, we identify and prove that Lasso actually behaves similarly as ℓ_1-LogR and RISE in the whole paramagnetic phase (as opposed to the limit regime $\theta^*_{r,t} \to 0, \forall (r,t) \in E$). We hope that our study could inspire further research on alternative simple and efficient methods for Ising model selection.

1.2 Related Works

In [Bento and Montanari, 2009], the authors pointed out a potential relevance of the incoherence condition of Lasso [Zhao and Yu, 2006] to ℓ_1-LogR by expanding the logistic loss around the true interactions θ^*. However, on the one hand, it is restricted to the case when the ℓ_1 regularization parameter approaches zero. On the other hand, the resultant quadratic loss is actually different from that of Lasso. Later, [Lokhov et al., 2018] observed that at high temperatures when the magnitude of interactions approaches zero, i.e., $\theta^*_{r,t} \to 0, \forall (r,t) \in E$, both the logistic and interaction screening objective (ISO) losses can be approximated as a square loss using a second-order Taylor expansion around zero (as opposed to θ^* in [Zhao and Yu, 2006]). However, their results are severely limited to the regime $\theta^*_{r,t} \to 0, \forall (r,t) \in E$. In other words, [Lokhov et al., 2018] attributed the potential success of Lasso to its similarity with ℓ_1-LogR/RISE in the regime $\theta^*_{r,t} \to 0, \forall (r,t) \in E$. Moreover, without considering the ℓ_1 regularization term, [Lokhov et al., 2018] only compared the analytical solution with that of the naive mean-field method [Tanaka, 1998, Kappen and Rodríguez, 1998, Ricci-Tersenghi, 2012]. A rigorous theoretical analysis of the consistency of Lasso for Ising model selection is still lacking.

[1] In fact, even for classification, linear regression is widely used, e.g., ridge classification [Dobriban and Wager, 2018], which can be significantly faster than logistic regression with a high number of classes [Scikit-learn,].

To the best of our knowledge, the first explicit analysis of Lasso for Ising model selection is given in [Meng et al., 2021] using statistical physics methods, building on previous studies [Bachschmid-Romano and Opper, 2017, Abbara et al., 2020, Meng et al., 2020]. In particular, [Meng et al., 2021] demonstrated that Lasso has the same order of sample complexity as ℓ_1-LogR for random regular (RR) graphs in the paramagnetic phase [Mezard and Montanari, 2009]. Furthermore, [Meng et al., 2021] provided an accurate estimate of the typical sample complexity as well as a precise prediction of the non-asymptotic learning performance. However, there are several limitations in [Meng et al., 2021]. First, since the replica method [Opper and Saad, 2001, Nishimori, 2001, Mezard and Montanari, 2009] they use is a non-rigorous method from statistical mechanics, a rigorous mathematical proof has remained lacking. Second, the results in [Meng et al., 2021] are restricted to the special class of RR graphs. In addition, since their analysis relies on the *self averaging property* [Nishimori, 2001, Mezard and Montanari, 2009], the results in [Meng et al., 2021] are meaningful in terms of the "typical case" [Engel and Van den Broeck, 2001] rather than the worst case. Moreover, [Meng et al., 2021] did not analyze the case of Lasso with post-thresholding.

Regarding the study of Lasso for nonlinear (not necessarily binary) targets, the past few years have seen an active line of research in the field of signal processing with a special focus on the single-index model [Brillinger, 1982, Plan and Vershynin, 2016, Thrampoulidis et al., 2015, Zhang et al., 2016, Genzel, 2016]. These studies are related to ours but with several important differences. First, in our study, the covariates are generated from an Ising model rather than a Gaussian distribution. Second, we focus on model selection consistency of Lasso while most previous studies considered estimation consistency except [Zhang et al., 2016]. However, [Zhang et al., 2016] only considered the classical asymptotic regime while we are interested in the high-dimensional setting where $n \ll p$. Another closely related work is [Erdogdu et al., 2019], which studied the relationship between the true minimizer of the population risk of a generalized linear model and the ordinary least square coefficient. Nevertheless, they only focused on the classic $n \gg p$ case. Moreover, even in the classic case, [Erdogdu et al., 2019] did not provide a rigorous analysis of the model selection consistency of Lasso with the empirical risk.

1.3 Notations

For each vertex $r \in V$, the neighborhood set is denoted as $\mathcal{N}(r) \coloneqq \{t \in V | (r,t) \in E\}$, the signed neighborhood set is defined as $\mathcal{N}_{\pm}(r) \coloneqq \{\text{sign}(\theta^*_{rt})\, t | t \in \mathcal{N}(r)\}$, and the corresponding node degree is denoted as $d_r \coloneqq |\mathcal{N}(r)|$. The maximum node degree of the whole graph G is denoted as $d \coloneqq \max_{r \in V} d_r$. We use $\mathcal{G}_{p,d}$ to denote the ensemble of graphs G with p vertices and maximum (not necessarily bounded) node degree $d \geq 3$. The minimum and maximum magnitudes of the interactions θ^*_{rt} for $(r,t) \in E$ are respectively denoted as

$$\theta^*_{\min} \coloneqq \min_{(r,t)\in E} |\theta^*_{rt}|\,, \quad \theta^*_{\max} \coloneqq \max_{(r,t)\in E} |\theta^*_{rt}|\,. \tag{2}$$

$\mathbb{E}_{\theta^*}\{\cdot\}$ denotes expectation w.r.t. the joint distribution $\mathbb{P}_{\theta^*}(x)$ (1). $|||A|||_\infty = \max_j \sum_k |A_{jk}|$ is the ℓ_∞ matrix norm of a matrix A. $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ denote the minimum and maximum eigenvalue of A, respectively.

2 Problem Setup

The problem of Ising model selection can be generally described as follows: given a collection of n i.i.d. samples $\mathfrak{X}_n \coloneqq \{x^{(1)}, \ldots, x^{(n)}\}$ from an Ising model defined on a graph $G = (V, E)$, the goal is to reconstruct the graph structure of G. In this paper we focus on Ising models defined on general locally tree-like graphs, i.e., the local neighborhood of a uniformly random vertex of the graph converges in distribution to a random rooted tree [Dembo and Montanari, 2010]. In particular, we also pay a special attention to the popular random regular (RR) graphs, one typical class of locally tree-like graphs with regular node degree $d_r = d$ and uniform couplings $\theta^*_{r,t} = \theta_0, \forall (r,t) \in E$.

As in [Ravikumar et al., 2010], we consider the slightly stronger criterion of *signed edge* recovery, and investigate the sufficient conditions on the *sparsistency property*.

Definition 1. *(signed edge) The signed edge set E^* of one Ising model with interactions θ^* is defined as $E^* \coloneqq \{\text{sign}(\theta^*_{rt})\}$ where sign $(\cdot)$ is an element-wise operation that maps every positive entry to 1, negative entry to -1, and zero entry to zero.*

Definition 2. *(sparsistency property) Suppose that $\hat{E}_n$ is an estimator of the signed edge E^* given $\mathfrak{X}_n$, then it is called (signed) model selection consistent in the sense that*

$$\mathbb{P}\left(\hat{E}_n = E^*\right) \to 1 \text{ as } n \to +\infty, \tag{3}$$

which is known as the sparsistency property [Ravikumar et al., 2010].

Our goal is to investigate the *sparsistency* property of Lasso [Tibshirani, 1996] for high-dimensional Ising models on locally tree-like graphs. Since recovering the edge set E^* of any graph $G = (V, E)$ is equivalent to reconstructing the associated signed neighborhood set $\mathcal{N}_{\pm}(r) \coloneqq \{\text{sign}(\theta^*_{rt})\, t | t \in \mathcal{N}(r)\}$ for each vertex $r \in V$ [Ravikumar et al., 2010], one can equivalently investigate the scaling condition on (n, p, d) which ensures that the

estimated signed neighborhood $\hat{\mathcal{N}}_{\pm}(r)$ agrees with the true neighborhood, i.e., $\left\{\hat{\mathcal{N}}_{\pm}(r) = \mathcal{N}_{\pm}(r), \forall r \in V\right\}$, with high probability.

Specifically, the estimate of the sub-vector $\theta^*_{\backslash r} := \{\theta^*_{rt} | t \in V \setminus r\} \in \mathbb{R}^{p-1},\ \forall r \in V$ is obtained via Lasso as follows

$$\hat{\theta}_{\backslash r} = \arg\min_{\theta_{\backslash r}} \left\{ \ell\left(\theta_{\backslash r}; \mathfrak{X}_n\right) + \lambda_{(n,p,d)} \left\|\theta_{\backslash r}\right\|_1 \right\}, \quad (4)$$

where $\ell\left(\theta_{\backslash r}; \mathfrak{X}_n\right)$ denotes the square loss function

$$\ell\left(\theta_{\backslash r}; \mathfrak{X}_n\right) := \frac{1}{2n} \sum_{i=1}^{n} \left(x_r^{(i)} - \sum_{u \in V \backslash r} \theta_{ru} x_u^{(i)}\right)^2, \quad (5)$$

and $\lambda_{(n,p,d)} > 0$ is the regularization parameter. For simplicity, instead of $\lambda_{(n,p,d)}$, λ_n will be used hereafter.

Subsequently, one can obtain an estimate $\hat{\mathcal{N}}_{\pm}(r)$ of $\mathcal{N}_{\pm}(r)$ from the Lasso results $\hat{\theta}_{\backslash r}$ in (4). Here we focus on two different settings: *without post-theresholding* and *with post-thresholding*. Without post-theresholding, one can simply estimate $\hat{\mathcal{N}}_{\pm}(r)$ using the sign information as [Ravikumar et al., 2010]

$$\hat{\mathcal{N}}_{\pm}(r) := \left\{ \operatorname{sign}\left(\hat{\theta}_{rt}\right) t | t \in V \setminus r, \hat{\theta}_{rt} \neq 0 \right\}. \quad (6)$$

Alternatively, one introduces a threshold $\xi > 0$ and then perform post-thresholding on $\hat{\theta}_{\backslash r}$ [Ekeberg et al., 2013, Decelle and Ricci-Tersenghi, 2014, Lokhov et al., 2018], leading to

$$\hat{\mathcal{N}}_{\pm}(r) := \left\{ \operatorname{sign}\left(\hat{\theta}_{rt}\right) \mathtt{1}\left(\left|\hat{\theta}_{rt}\right| > \xi\right) t | t \in V \setminus r, \hat{\theta}_{rt} \neq 0 \right\}, \quad (7)$$

where $\mathtt{1}(\cdot)$ is an indicator function that equals to 1 if the event is true and 0 otherwise.

3 Main results

3.1 Preliminary Results

Before stating the main results, we first present two different results of Lasso compared with ℓ_1-LogR regarding the expected first and second derivative of the loss function, i.e., $\mathbb{E}_{\theta^*}\{\nabla \ell\left(\theta_{\backslash r}; \mathfrak{X}_1^n\right)\}$ and $\mathbb{E}_{\theta^*}\{\nabla^2 \ell\left(\theta_{\backslash r}; \mathfrak{X}_1^n\right)\}$.

Lemma 1. *For general tree-like graphs in the paramagnetic phase, the solution to $\mathbb{E}_{\theta^*}\{\nabla \ell\left(\theta_{\backslash r}; \mathfrak{X}_1^n\right)\} = 0$, denoted as $\tilde{\theta}^*_{\backslash r} = \left\{\tilde{\theta}^*_{rt}\right\}_{t \in V \backslash r} \in \mathbb{R}^{p-1}$, can be obtained as*

$$\tilde{\theta}^*_{rt} = \begin{cases} \frac{\tanh(\theta^*_{rt}) / \left(1 - \tanh^2(\theta^*_{rt})\right)}{1 - d_r + \sum_{u \in \mathcal{N}(r)} \frac{1}{1 - \tanh^2(\theta^*_{ru})}} & \text{if } (r,t) \in E \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

where d_r is the node degree of r. In particular, for RR graph with uniform coupling strength $\theta^*_{rt} = \theta_0, \forall (r,t) \in E$ and constant node degree $d_r = d$, there is

$$\tilde{\theta}^*_{rt} = \begin{cases} \frac{\tanh(\theta_0)}{1 + (d-1)\tanh^2(\theta_0)} & \text{if } (r,t) \in E \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

Proof. See Appendix A. □

Lemma 1 indicates that, the solution $\tilde{\theta}^*_{\backslash r}$ is a rescaled value of the true parameter $\theta^*_{\backslash r}$ and thus shares the same sign structure, i.e., $\operatorname{sign}\left(\tilde{\theta}^*_{\backslash r}\right) = \operatorname{sign}\left(\theta^*_{\backslash r}\right)$. The minimum magnitude of $\tilde{\theta}^*_{rt}$ for $(r,t) \in E$ in (8) is denoted as

$$\tilde{\theta}^*_{\min} := \min_{(r,t) \in E} \tilde{\theta}^*_{rt}. \quad (10)$$

For the second derivative or Hessian matrix, in the case of Lasso, it corresponds exactly to the covariance matrix, i.e.,

$$Q^*_r := \mathbb{E}_{\theta^*}\{\nabla^2 \ell\left(\theta_{\backslash r}; \mathfrak{X}_1^n\right)\} = \mathbb{E}_{\theta^*}\{X_{\backslash r} X_{\backslash r}^T\}, \forall r \in V. \quad (11)$$

As opposed to [Ravikumar et al., 2010], the additional variance function term of ℓ_1-LogR (eq. (12) in [Ravikumar et al., 2010], denoted as $\eta(X; \theta^*)$) does not exist in Q^*_r (11), which makes Lasso different from ℓ_1-LogR, including its behavior and the corresponding proof. For notational simplicity, Q^*_r will be written as Q^* hereafter. Denote $S := \{(r,t) \mid t \in \mathcal{N}(r)\}$ as the subset of indices associated with edges of r and S^c as its complement. The $d_r \times d_r$ sub-matrix of Q^* indexed by S is denoted as Q^*_{SS}. Other sub-matrices like $Q^*_{S^cS}$ are defined in the same way.

3.2 Lasso without Post-thresholding

For Lasso without post-thresholding, i.e., the signed edge set $\hat{\mathcal{N}}_{\pm}(r), \forall r \in V$ is obtained as (6), we have

Theorem 1. *(RR graphs) Consider a collection of n i.i.d. samples $\mathfrak{X}_n := \left\{x^{(1)}, \ldots, x^{(n)}\right\}$ drawn from an Ising model on a RR graph $G = (V, E) \in \mathcal{G}_{p,d}$ with regular node degree d and uniform couplings $\theta^*_{r,t} = \theta_0, \forall (r,t) \in E$. Suppose that the Ising model is in the paramagnetic phase, i.e., $(d-1)\tanh\theta_0 < 1$, then there exist constants L, c independent of (n,p,d), so that the Lasso estimator (4) with the regularization parameter $\lambda_n \leq \frac{\tanh(\theta_0)(1-\tanh^2(\theta_0))}{6\sqrt{d}(1+(d-1)\tanh^2(\theta_0))}$ reconstructs the signed edge set by (6) perfectly with probability at least*

$$\mathbb{P}\left(\hat{E}_n = E^*\right) \geq 1 - 2\exp\left(-c\lambda_n^2 n\right) \quad (12)$$

as long as $n \geq \max\left\{Ld^3, \frac{64(1+\tanh(\theta_0))^2}{(1-\tanh(\theta_0))^2 \lambda_n^2}\right\} \log p$.

Remark 3: Theorem 1 indicates that the probability that the Lasso estimator (4) successfully recovers the true signed edge set decays exponentially as a function of $\lambda_n^2 n$, which is the same as ℓ_1-LogR [Ravikumar et al., 2010]. If λ_n is chosen such that $\lambda_n^2 n \to \infty$ as $n \to \infty$, Lasso is model

selection consistent, i.e., $\mathbb{P}\left(\hat{E}_n = E^*\right) \to 1$ as $n \to \infty$. In the high-dimensional case, one reasonable choice of λ_n that satisfies both Theorem 1 and $\lambda_n^2 n \to \infty$ is $\lambda_n = \kappa\sqrt{\frac{\log p}{n}}$, where $\kappa \geq \frac{8(1+\tanh\theta_0)}{1-\tanh\theta_0}$. In this case, i.e., $\lambda_n = \kappa\sqrt{\frac{\log p}{n}}$, from Theorem 1, it is obtained that the number of samples required for model selection consistency needs to satisfy $n \geq \max\left\{Ld^3, \frac{36\kappa^2(1+(d-1)\tanh(\theta_0))^2}{\tanh^2(\theta_0))(1-\tanh^2(\theta_0))^2}d\right\}\log p$. In practical applications, one might not be able to obtain the ideal κ when θ_0 is unknown. However, a larger κ can be chosen, which is possible either by using a prior knowledge of the range of θ_0 for a specific problem, or by setting κ as large as possible. A disadvantage of larger κ is that, the number of samples n becomes larger since $n \geq \max\left\{Ld^3, \frac{36\kappa^2(1+(d-1)\tanh(\theta_0))^2}{\tanh^2(\theta_0))(1-\tanh^2(\theta_0))^2}d\right\}\log p$. However, this is an inevitable price we have to pay due to the lack of knowledge of the models.

Note that while the uniform coupling of RR graphs in Theorem 1 is a limitation, Theorem 1 holds without additional assumptions as ℓ_1-LogR [Ravikumar et al., 2010]. For general locally tree-like graphs, under additional mild assumptions similar to ℓ_1-LogR [Ravikumar et al., 2010], namely the *dependency condition* and *incoherence condition*, we can still obtain similar results as RR graphs in Theorem 1.

Condition 1 (C1): *dependency condition*. The sub-matrix Q^*_{SS} has bounded eigenvalue, i.e., there exists a constant $C_{\min} > 0$ such that

$$\Lambda_{\min}\left(Q^*_{SS}\right) \geq C_{\min}. \tag{13}$$

Condition 2 (C2): *incoherence condition*. There exists an $\alpha \in (0, 1]$ such that

$$\left|\left|\left|Q^*_{S^cS}\left(Q^*_{SS}\right)^{-1}\right|\right|\right|_\infty \leq 1 - \alpha. \tag{14}$$

Theorem 2. *(tree-like graphs) Consider general tree-like graphs $G = (V, E) \in \mathcal{G}_{p,d}$ in the paramagnetic phase. Suppose that conditions (C1) and (C2) are satisfied by the population covariance matrix Q^*. If the regularization parameter λ_n is selected to satisfy $\lambda_n \geq \frac{4\sqrt{c+1}(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$ for some constant $c > 0$, then there exists a constant L independent of (n, p, d) such that if*

$$n \geq Ld^3 \log p, \tag{15}$$

then with probability at least $1 - 2\exp(-c\log p) \to 1$ as $p \to \infty$, the following properties hold:

(a) For each node $r \in V$, the Lasso estimator (4) has a unique solution, and thus uniquely specifies a signed neighborhood $\hat{\mathcal{N}}_\pm(r)$ with (6).

(b) For each node $r \in V$, the estimated signed neighborhood vector $\hat{\mathcal{N}}_\pm(r)$ with (6) correctly excludes all edges not in the true neighborhood. Moreover, it correctly includes all edges if the minimum magnitude of the rescaled parameter satisfies $\tilde{\theta}^*_{\min} \geq \frac{6\lambda_n\sqrt{d}}{C_{\min}}$.

Remark 4: Theorem 2 indicates that the probability that Lasso recovers the true signed edge set $\mathbb{P}\left(\hat{E}_n = E^*\right) \to 1$ exponentially as a function of $\log p$. Hence, for tree-like Ising models in the paramagnetic phase, under conditions (C1) and (C2), in the high-dimensional setting (for $p \to \infty$), Lasso is model selection consistent with $n = \Omega(d^3 \log p)$ samples, which is the same as ℓ_1-LogR [Ravikumar et al., 2010].

In contrast to RR graphs in Theorem 1, for general tree-like graphs, two additional assumptions (C1) and (C2) are imposed for the success of Lasso without post-thresholding. However, it is worth noting that ℓ_1-LogR without post-thresholding also suffers from the same limitation as shown in [Ravikumar et al., 2010], which is due to the fundamental difficulty in verifying (C1) and (C2) for general graphs.

3.3 Lasso with Post-thresholding

For Lasso with post-thresholding, i.e., the signed neighborhood set $\hat{\mathcal{N}}_\pm(r), \forall r \in V$ is obtained as (7), we obtain the following results.

Theorem 3. *(Square error, tree-like graphs) Consider an Ising model defined on tree-like graphs $G = (V, E) \in \mathcal{G}_{p,d}$. $\forall r \in V$ and for any $\varepsilon_1 > 0$, in the paramagnetic phase, the square error of the Lasso estimator (4) with regularization parameter $\lambda_n = 4\sqrt{\frac{\log\frac{3p}{\varepsilon_1}}{n}}$ is bounded with probability at least $1 - \varepsilon_1$ by*

$$\left\|\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}\right\|_2 \leq 2^6\sqrt{d}\,(d+1)\,e^{2\theta^*_{\max}d}\sqrt{\frac{\log\frac{3p}{\varepsilon_1}}{n}} \tag{16}$$

when $n \geq 2^{14}d^2\,(d+1)^2\,e^{4\theta^*_{\max}d}\log\frac{3p^2}{\varepsilon_1}$.

Theorem 4. *(Structure learning, tree-like graphs) Consider an Ising model defined on tree-like graphs $G = (V, E) \in \mathcal{G}_{p,d}$. In the paramagnetic phase, for any $\varepsilon_2 > 0$, the Lasso estimator (4) with regularization parameter $\lambda_n = 4\sqrt{\frac{\log\frac{3p^2}{\varepsilon_2}}{n}}$ reconstructs the sign edge set by (7) perfectly with probability*

$$\mathbb{P}\left(\hat{E} = E^*\right) \geq 1 - \varepsilon_2, \tag{17}$$

as long as

$$n \geq \max\left\{d, \left(\tilde{\theta}^*_{\min}\right)^{-2}\right\} 2^{14}d\,(d+1)^2\,e^{4\theta^*_{\max}d}\log\frac{3p^3}{\varepsilon_2}. \tag{18}$$

Remark 5: Results in Theorems 3 and 4 hold for general tree-like graphs without any further assumptions of (C1) and (C2). In particular, Theorem 4 indicates that Lasso

with post-thresholding is model selection consistent under similar conditions as the RISE [Vuffray et al., 2016] and ℓ_1-LogR with post-thresholding [Lokhov et al., 2018]. Note that similarly as RISE and ℓ_1-LogR [Vuffray et al., 2016, Lokhov et al., 2018], the obtained bound in (18) is a rather loose bound, especially in the paramagnetic phase, e.g., while it suggests an exponential growth w.r.t. $\theta^*_{\max}$, it is in fact not the case in the paramagnetic phase (see Figure 4 in [Lokhov et al., 2018]).

Remark 6: In Theorems 3 and 4, the regularization parameter λ_n is chosen as a function of the controlled probability values ϵ_1 or ϵ_2. In practical applications when no prior knowledge is available, similarly as [Vuffray et al., 2016, Lokhov et al., 2018], we can simply choose λ_n as $\lambda_n = 4\sqrt{\frac{\log \frac{3p}{\epsilon_1}}{n}}$ or $\lambda_n = 4\sqrt{\frac{\log \frac{3p^2}{\epsilon_2}}{n}}$, respectively. Moreover, as described in [Lokhov et al., 2018], given a sufficient number of samples, other techniques such as consistency cross-validation can be used for selecting the optimal value of λ_n on a case-by-case basis. For more details, please refer to the supplementary material of [Lokhov et al., 2018].

4 Proof of the main results

Here we provide a sketch of the proofs for the main results. For details, please refer to Appendices D and E.

4.1 Sketch of the proof for Theorems 1 and 2

For the proof of Lasso without post-thresholding, we use the primal-dual witness proof framework [Ravikumar et al., 2010], which was originally proposed in [Wainwright, 2009]. The main idea of the primal-dual witness method is to explicitly construct an optimal primal-dual pair which satisfies the sub-gradient optimality conditions associated with the Lasso estimator (4). Subsequently, it is proved that under the stated assumptions on (n, p, d), the optimal primal-dual pair can be constructed such that they act as a witness, i.e., a certificate that guarantees that the neighborhood-based Lasso estimator (4) together with (6) correctly recovers the signed edge set of the graph $G \in \mathcal{G}_{p,d}$.

Generally speaking, the proof of Theorems 1 and 2 consists of two stages. At the first stage, we consider a "fixed design" case assuming that the sample Hessian $Q^n := \frac{1}{n}\sum_{i=1}^n x^{(i)}_{\backslash r}\left(x^{(i)}_{\backslash r}\right)^T$, satisfies both conditions (C1) and (C2). Afterwards, at the second stage, using some large-deviation analysis we provide guarantees under which both conditions (C1) and (C2) hold for the sample Hessian Q^n with high probability. Finally, we obtain Theorems 1 and 2 combining results of the two stages. Notably, for RR graphs, there is one remarkable property, as shown in Lemma 2:

Lemma 2. *For Ising models defined on RR graphs $G = (V, E) \in \mathcal{G}_{p,d}$ with regular node degree d and uniform couplings $\theta^*_{r,t} = \theta_0, \forall (r,t) \in E$. In the paramagnetic phase, both conditions (C1) and (C2) hold for Q^*, where $C_{\min} = 1 - \tanh^2\theta_0$ and $\alpha = 1 - \tanh\theta_0$.*

Proof. See Appendix B. □

As a result, in Theorem 1, there is no need for assumptions (C1) and (C2) in the case of RR graphs.

The important results at the first stage are shown in Proposition 1 and Proposition 2, which correspond to the RR graphs and general tree-like graphs, respectively.

Proposition 1. *(fixed design, RR graphs) Consider an Ising model on a RR graph $G = (V, E) \in \mathcal{G}_{p,d}$ with regular node degree d and uniform couplings $\theta^*_{r,t} = \theta_0, \forall (r,t) \in E$. Suppose that the Ising model is in the paramagnetic phase, and that the sample Hessian Q^n satisfies (C1) and (C2). If the regularization parameter λ_n satisfies $\lambda_n \geq \frac{8(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$, then with probability at least $1 - 2\exp\left(-c\lambda_n^2 n\right) \to 1$, the following properties hold:*

(a) For each node $r \in V$, the Lasso estimator (4) has a unique solution, and thus uniquely specifies a signed neighborhood $\hat{\mathcal{N}}_{\pm}(r)$.

(b) For each node $r \in V$, the estimated signed neighborhood vector $\hat{\mathcal{N}}_{\pm}(r)$ using the Lasso estimator (4) correctly excludes all edges not in the true neighborhood. Moreover, it correctly includes all edges if $\tilde{\theta}^*_{\min} \geq \frac{6\lambda_n\sqrt{d}}{C_{\min}}$.

Proof. See Appendix D.1. □

Proposition 2. *(fixed design, tree-like graphs) Consider an Ising model defined on a tree-like graph $G = (V, E) \in \mathcal{G}_{p,d}$ with parameter vector θ^* and associated signed edge set E^*. Suppose that the Ising model is in the paramagnetic phase, and the sample Hessian Q^n satisfies (C1) and (C2) and the regularization parameter λ_n satisfies $\lambda_n \geq \frac{4\sqrt{c+1}(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$ for some constant $c > 0$. Under these conditions, if*

$$n \geq (c+1)d^2 \log p, \tag{19}$$

then with probability at least $1 - 2\exp(-c\log p) \to 1$ as $p \to \infty$, the following properties hold:

(a) For each node $r \in V$, the Lasso estimator (4) has a unique solution, and thus uniquely specifies a signed neighborhood $\hat{\mathcal{N}}_{\pm}(r)$.

(b) For each node $r \in V$, the estimated signed neighborhood vector $\hat{\mathcal{N}}_{\pm}(r)$ correctly excludes all edges not in the true neighborhood. Moreover, it correctly includes all edges if $\tilde{\theta}^*_{\min} \geq \frac{6\lambda_n\sqrt{d}}{C_{\min}}$, where $\tilde{\theta}^*_{\min}$ is the minimum magnitude of the rescaled parameter $\tilde{\theta}^*$ defined in (8).

Proof. See Appendix D.2. □

Note that in the above two Propositions of the"fixed design" case, in contrast to the "fixed design" results of ℓ_1-LogR in [Ravikumar et al., 2010], there is no requirement of an additional scaling condition of $n \geq Ld^2 \log p$. This is due to the fundamental difference between the square loss of Lasso and the logistic loss of ℓ_1-LogR. Specifically for ℓ_1-LogR, $n \geq Ld^2 \log p$ is needed to ensure the ℓ_2-consistency of the primal sub-vector and to bound the remainder term, while it is not the case for Lasso with square loss, as shown in Lemma 5. However, this only holds under the assumption that the sample Hessian satisfies conditions (C1) and (C2). To ensure that these conditions are satisfied by the sample Hessian, an additional requirement of $n \geq Ld^3 \log p$ is still needed, as shown in the final results in Theorem 1 and Theorem 2.

Some key results: The key results for the proofs of Lasso without post-thresholding are given as follows.

Lemma 3. *Denote* $W^n = -\nabla \ell\left(\tilde{\theta}^*_{\backslash r}; \mathfrak{X}_1^n\right)$. *The s-th element of* W^n, *denoted as* W_s^n, *can be written as follows*

$$W_s^n = \frac{1}{n}\sum_{i=1}^{n} Z_s^{(i)},\ \forall s \in V \setminus r, \tag{20}$$

$$Z_s^{(i)} \coloneqq x_s^{(i)}(x_r^{(i)} - \sum_{t \in V\setminus r} \tilde{\theta}^*_{rt} x_t^{(i)}). \tag{21}$$

Then, $\mathbb{E}_{\theta^*}\left(Z_s^{(i)}\right) = 0$, $\texttt{Var}\left(Z_s^{(i)}\right) \leq 1$. *Furthermore:*

(a) For RR graphs, there is $\left|Z_s^{(i)}\right| \leq 2$;

(b) For general tree-like graphs, there is $\left|Z_s^{(i)}\right| \leq d$.

Proof. See Appendix C.1. □

The behavior of $\|W^n\|_\infty$ is shown in Lemma 4.

Lemma 4. *Regarding* $W^n = -\nabla \ell\left(\tilde{\theta}^*_{\backslash r}; \mathfrak{X}_n\right)$ *in Lemma 3:*

(a) For RR graphs, if $\lambda_n \geq \frac{8(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$, *then*

$$\mathbb{P}\Big(\frac{2-\alpha}{\lambda_n}\|W^n\|_\infty \geq \frac{\alpha}{2}\Big) \leq 2\exp\Big(-\frac{\alpha^2\lambda_n^2 n}{32(2-\alpha)^2} + \log p\Big), \tag{22}$$

(b) For general tree-like graphs, if $n \geq (c+1)\, d^2 \log p$ *for some constant* $c > 0$ *and* $\lambda_n \geq \frac{4\sqrt{c+1}(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$, *then*

$$\mathbb{P}\Big(\frac{2-\alpha}{\lambda_n}\|W^n\|_\infty \geq \frac{\alpha}{2}\Big) \leq 2\exp\left(-c\log p\right). \tag{23}$$

Proof. See Appendix C.2. □

Lemma 5. *If* $\|W^n\|_\infty \leq \frac{\lambda_n}{2}$, *then there is*

$$\left\|\hat{\theta}_S - \tilde{\theta}^*_S\right\|_2 \leq \frac{3}{C_{\min}}\lambda_n\sqrt{d}. \tag{24}$$

Proof. See Appendix C.3. □

4.2 Sketch of the proof for Theorems 3 and 4

In proving Theorems 3 and 4, we resort to the restricted strong convexity framework in [Negahban et al., 2012].

First, consider the proof of Theorem 3 which provides an estimation error bound (16) of Lasso. Similarly as RISE and ℓ_1-LogR [Vuffray et al., 2016, Negahban et al., 2012, Lokhov et al., 2018], to obtain a handle on the (rescaled) square error of Lasso, two sufficient conditions (C3) and (C4) are enforced as follows:

Condition 3 (C3): The ℓ_1 regularization parameter λ_n strongly enforces regularization if it is greater than any partial derivatives of the loss function $\ell\left(\theta_{\backslash r}; \mathfrak{X}_1^n\right)$ evaluated at $\tilde{\theta}^*_{\backslash r}$ defined in (8), i.e.,

$$\lambda_n \geq 2\left\|\nabla \ell\left(\tilde{\theta}^*_{\backslash r}; \mathfrak{X}_1^n\right)\right\|_\infty. \tag{25}$$

Condition (C3) guarantees that if the vector of the rescaled couplings $\tilde{\theta}^*_{\backslash r}$ has at most d non-zero elements, then the estimation difference $\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}$ lies within the set

$$K \coloneqq \left\{\triangle \in \mathbb{R}^{p-1} \mid \|\triangle\|_1 \leq 4\sqrt{d}\,\|\triangle\|_2\right\}. \tag{26}$$

Condition 4 (C4): The square loss of Lasso is restricted strongly convex w.r.t. set K (26) on a ball of radius R centered at $\theta_{\backslash r} = \tilde{\theta}^*_{\backslash r}$ if for all $\triangle_{\theta_{\backslash r}} \in K$ such that $\left\|\triangle_{\theta_{\backslash r}}\right\|_2 \leq R$, there exists a constant $\kappa > 0$ such that the remainder of the first-order Taylor expansion of the loss function satisfies

$$\delta\ell\left(\triangle_{\theta_{\backslash r}}, \tilde{\theta}^*_{\backslash r}; \mathfrak{X}_1^n\right) \geq \kappa\left\|\triangle_{\theta_{\backslash r}}\right\|_2^2. \tag{27}$$

where $\triangle_{\theta_{\backslash r}} \in \mathbb{R}^{p-1}$ is an arbitrary vector and the remainder can be calculated as

$$\delta\ell\left(\triangle_{\theta_{\backslash r}}, \tilde{\theta}^*_{\backslash r}; \mathfrak{X}_1^n\right) = \frac{1}{2}\triangle^T_{\theta_{\backslash r}} Q^n \triangle_{\theta_{\backslash r}}. \tag{28}$$

The key point is that, the estimation error $\left\|\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}\right\|_2$ of Lasso can be controlled if conditions (C3) and (C4) are satisfied, as shown in Proposition 3:

Proposition 3. *(Theorem 1, [Negahban et al., 2012]) If the Lasso estimator (4) satisfies both (C3) and (C4) with* $R \geq 3\sqrt{d}\frac{\lambda_n}{\kappa}$, *then the square error is bounded by*

$$\left\|\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}\right\|_2 \leq 3\sqrt{d}\frac{\lambda_n}{\kappa}. \tag{29}$$

As a result, the proof of Theorem 3 is done through Proposition 3 by evaluating the two conditions (C3) and (C4).

Regarding the proof of Theorem 4, it is simply an application of Theorem 3 by choosing a specific value of the estimation error. Specifically, with the definition of $\tilde{\theta}^*_{\min}$ in

(8) as the minimum rescaled coupling for a general graph, suppose that the estimated error $\left\|\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}\right\|_2$ is controlled to be smaller than $\tilde{\theta}^*_{\min}/2$, then one can readily recover the structure of the neighborhood of node r by setting the edges whose absolute estimated couplings are less than $\tilde{\theta}^*_{\min}/2$ to be absent [Lokhov et al., 2018]. Subsequently, repeating this procedure over all the p vertices, we are guaranteed through the union bound that exact reconstruction of the full edge set E^* can be obtained with some predefined probability.

Some key results: The key results for the proofs of Lasso with post-thresholding are given as follows. Specifically, Lemma 7 and Lemma 8 are used to prove Lemma 9, which is then combined with Lemma 6 to evaluate the conditions (C3) and (C4) via Proposition 3, leading to the proof of Theorem 3.

Lemma 6. *For any $\varepsilon_3 > 0$, if $n \geq d^2 \log \frac{2p}{\varepsilon_3}$, then probability at least $1 - \varepsilon_3$*

$$\|W^n\|_\infty \leq 2\sqrt{\frac{\log \frac{2p}{\varepsilon_3}}{n}}. \tag{30}$$

Proof. See Appendix C.4. □

The randomness of $\delta\ell\left(\triangle_{\theta_{\backslash r}}, \tilde{\theta}^*_{\backslash r}; \mathfrak{X}_1^n\right)$ can be controlled by Q^n, which concentrates towards its mean independently of $\triangle_{\theta_{\backslash r}}$, as shown in following lemma

Lemma 7. *Let $\epsilon > 0$, $\varepsilon_4 > 0$ and $n \geq \frac{2}{\epsilon^2} \log \frac{p^2}{\varepsilon_4}$, then with probability greater than $1 - \varepsilon_4$, we have for all $s, t \in V \setminus r$*

$$|Q^n_{st} - Q^*_{st}| \leq \epsilon,$$

where $Q^n_{st} = \frac{1}{n}\sum_{i=1}^n x_t^{(i)} x_t^{(i)}$ and $Q^*_{st} = \mathbb{E}_{\theta^*}\left(x_s^{(i)} x_t^{(i)}\right), s, t \in V \setminus r$.

Proof. See Appendix C.5. □

Lemma 8 states that the smallest eigenvalue of Q^* is bounded below from zero independent of p.

Lemma 8. *(Lemma 7 in [Vuffray et al., 2016]) For Ising model with graph $G \in \mathcal{G}_{p,d}$ with maximum coupling strength $\theta^*_{\max}$. Then for all $\triangle_{\theta_{\backslash r}} \in \mathbb{R}^{p-1}$, we have*

$$\triangle^T_{\theta_{\backslash r}} Q^* \triangle_{\theta_{\backslash r}} \geq \frac{e^{-2\theta^*_{\max} d}}{d+1} \left\|\triangle_{\theta_{\backslash r}}\right\|_2^2.$$

Given the above results, the restricted strong convexity of the square loss (5) for Ising model problems is stated as follows.

Lemma 9. *For Ising model with graph $G \in \mathcal{G}_{p,d}$ with maximum coupling strength $\theta^*_{\max}$, $\forall \varepsilon_4 > 0$, when $n > 2^{11} d^2 (d+1)^2 e^{4\theta^*_{\max} d} \log \frac{p^2}{\varepsilon_4}$, the square loss (5) of Lasso satisfies, with probability at least $1 - \varepsilon_4$, the restricted strong convexity condition*

$$\delta\ell\left(\triangle_{\theta_{\backslash r}}, \tilde{\theta}^*_{\backslash r}; \mathfrak{X}_1^n\right) \geq \frac{e^{-2\theta^*_{\max} d}}{4(d+1)} \left\|\triangle_{\theta_{\backslash r}}\right\|_2^2 \tag{31}$$

for all $\triangle_{\theta_{\backslash r}} \in \mathbb{R}^{p-1}$ such that $\left\|\triangle_{\theta_{\backslash r}}\right\|_1 \leq 4\sqrt{d}\left\|\triangle_{\theta_{\backslash r}}\right\|_2$.

Proof. See Appendix C.6. □

5 Experimental Results

In this section we conduct simulations to verify our theoretical findings that, simply speaking, Lasso performs similarly as ℓ_1-LogR on typical tree-like graphs in the paramagnetic phase. Two different structures of tree-like graphs are evaluated, namely RR graphs and star-shaped graphs. In addition, to have a look at the performance of Lasso for graphs with many loops, we also evaluate the square lattice (grid) graphs with periodic boundary condition. It is worth noting that the RR and star-shaped graphs represent graphs with bounded node degree (the maximum node degree d is a fixed constant) and unbounded node degree (the maximum node degree d grows as the size of p), respectively.

The experimental procedures are as follows. First, a graph $G = (V, E) \in \mathcal{G}_{p,d}$ is generated and the Ising model is defined on it. Then, the spin snapshots are obtained using Monte-Carlo sampling, yielding the dataset $\mathfrak{X}_1^n$. The regularization parameter is set to be a constant factor of $\sqrt{\frac{\log p}{n}}$. For any graph, we performed simulations using neighborhood-based Lasso (4) $\forall r \in V$ and then the associated signed neighborhood $\hat{\mathcal{N}}_\pm(r)$ is estimated as (6). Similar to [Ravikumar et al., 2010], the sample size n scaling is set to be proportional to $d \log p$. For comparison, the results of the ℓ_1-LogR estimator [Ravikumar et al., 2010] are also shown. The results are averaged over 200 trials in all cases.

The results of RR graph and grid graph are shown in Figure 1. In both cases, even for grid graph with many loops, using the Lasso estimator, all curves for different model sizes p line up with each other well, demonstrating that for a graph with fixed degree d, the ratio $n/\log p$ controls the success or failure of the Ising model selection. Importantly, the behavior of Lasso is about the same as ℓ_1-LogR.

Figure 2 shows results for star-shaped graph whose maximum degree d is unbounded and grows as the dimension p grows. Two kinds of star-shaped graphs are considered by designating one node as the hub and connecting it to $d < (p-1)$ of its neighbors. Specifically, for linear sparsity, it is assumed that $d = \lceil 0.1p \rceil$ while for logarithmic sparsity, we assume $d = \lceil \log p \rceil$. We use positive interactions and set the active interactions to be $\theta^*_{rt} = \frac{1.2}{\sqrt{d}}$ for all $(r, t) \in E$ as [Ravikumar et al., 2010]. As depicted in

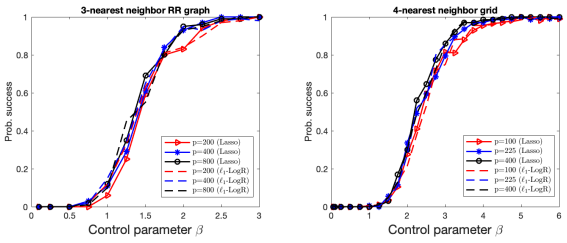

Figure 1: Success probability versus the control parameter β for Ising models. Left: RR graph with $d = 3$ and mixed interactions $\theta^*_{rt} = \pm 0.4$ for all $(r, t) \in E$, $\beta = \frac{n}{10d \log p}$; Right: 4-nearest neighbor grid graph with $d = 4$ and positive interactions $\theta^*_{rt} = 0.2$ for all $(r, t) \in E$, $\beta = \frac{n}{15d \log p}$.

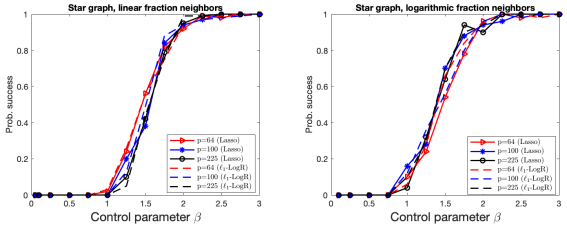

Figure 2: Success probability versus the control parameter $\beta = \frac{n}{10d \log p}$ for Ising models on star-shaped graphs for attractive interactions $\theta^*_{rt} = \frac{1.2}{\sqrt{d}}$ for all $(r, t) \in E$. Left: linear growth in degrees, i.e., $d = \lceil 0.1p \rceil$; Right: logarithmic growth in degrees, i.e., $d = \lceil \log p \rceil$.

Figure 2, Lasso behaves similarly as ℓ_1-LogR in both cases, which is consistent with our theoretical analysis.

6 Conclusion

We have theoretically analyzed the model selection consistency of Lasso, *both with and without post-thresholding*, for the problem of high-dimensional Ising model selection with a focus on the paramagnetic phase. Specifically, in the case without post-thresholding, we prove that Lasso is model selection consistent with the same order of sample complexity as that of ℓ_1-LogR for RR graphs. For general tree-like graphs, similar result is obtained under mild assumptions of the *dependency condition* and *incoherence condition*. Moreover, in the case with post-thresholding, for general tree-like graphs, we not only obtain an upper bound of the reconstructed square error of Lasso, but also prove the consistency of Lasso with post-thresholding with the same order of sample complexity as that of ℓ_1-LogR and RISE without any assumptions on the *dependency condition* and *incoherence condition*. Experimental results are consistent with the theoretical analysis.

There are several interesting future directions for current study. First, since our focus in this paper is the paramagnetic phase, one important future work is to extend the current analysis to high-dimensional Ising models defined on general graphs beyond the paramagnetic phase, e.g., ferromagnetic phase, to see whether it still can, similarly as ℓ_1-LogR and RISE, successfully recover the graph structure of Ising models with the same order of the number of samples. Another future work is to investigate the performance of Lasso for high-dimensional Ising model selection in the non-i.i.d. case [Dutt et al., 2021]. The study of other alternative simple and efficient methods for Ising model selection is also an interesting topic for future investigation.

Acknowledgements

This work was supported by JSPS KAKENHI Nos. 17H00764, 18K11463, and 19H01812, 22H05117, and JST CREST Grant Number JPMJCR1912, Japan.

References

[Abbara et al., 2020] Abbara, A., Kabashima, Y., Obuchi, T., and Xu, Y. (2020). Learning performance in inverse Ising problems with sparse teacher couplings. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(7):073402.

[Bachschmid-Romano and Opper, 2017] Bachschmid-Romano, L. and Opper, M. (2017). A statistical physics approach to learning curves for the inverse ising problem. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(6):063406.

[Bento and Montanari, 2009] Bento, J. and Montanari, A. (2009). Which graphical models are difficult to learn? In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 1303–1311.

[Besag, 1975] Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195.

[Bresler, 2015] Bresler, G. (2015). Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782.

[Brillinger, 1982] Brillinger, D. R. (1982). A generalized linear model with Gaussian regressor variables. In *A Festschrift for Erich L. Lehmann*, page 97–114.

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

[Dembo and Montanari, 2010] Dembo, A. and Montanari, A. (2010). Ising models on locally tree-like graphs. *The Annals of Applied Probability*, 20(2):565–592.

[Dobriban and Wager, 2018] Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.

[Dutt et al., 2021] Dutt, A., Lokhov, A. Y., Vuffray, M., and Misra, S. (2021). Exponential reduction in sample complexity with learning of Ising model dynamics. *arXiv preprint arXiv:2104.00995*.

[Ekeberg et al., 2013] Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707.

[Engel and Van den Broeck, 2001] Engel, A. and Van den Broeck, C. (2001). *Statistical mechanics of learning*. Cambridge University Press.

[Erdogdu et al., 2019] Erdogdu, M. A., Bayati, M., and Dicker, L. H. (2019). Scalable approximations for generalized linear problems. *The Journal of Machine Learning Research*, 20(1):231–275.

[Genzel, 2016] Genzel, M. (2016). High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. *IEEE Transactions on Information Theory*, 63(3):1601–1619.

[Gomila, 2021] Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, 150(4):700.

[Hoeffding, 1994] Hoeffding, W. (1994). Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer.

[Höfling and Tibshirani, 2009] Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10(4).

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

[Kappen and Rodríguez, 1998] Kappen, H. J. and Rodríguez, F. d. B. (1998). Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10(5):1137–1156.

[Koller and Friedman, 2009] Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

[Krishnan et al., 2020] Krishnan, J., Torabi, R., Schuppert, A., and Di Napoli, E. (2020). A modified Ising model of barabási–albert network with gene-type spins. *Journal of mathematical biology*, 81(3):769–798.

[Liebl and Zacharias, 2021] Liebl, K. and Zacharias, M. (2021). Accurate modeling of dna conformational flexibility by a multivariate Ising model. *Proceedings of the National Academy of Sciences*, 118(15).

[Lokhov et al., 2018] Lokhov, A. Y., Vuffray, M., Misra, S., and Chertkov, M. (2018). Optimal structure and parameter learning of Ising models. *Science advances*, 4(3):e1700791.

[Marbach et al., 2012] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804.

[McAuley and Leskovec, 2012] McAuley, J. J. and Leskovec, J. (2012). Learning to discover social circles in ego networks. volume 2012, pages 548–56. Citeseer.

[Meng et al., 2020] Meng, X., Obuchi, T., and Kabashima, Y. (2020). Structure learning in inverse Ising problems using ℓ_2-regularized linear estimator. *arXiv preprint arXiv:2008.08342*.

[Meng et al., 2021] Meng, X., Obuchi, T., and Kabashima, Y. (2021). Ising model selection using ℓ_1-regularized linear regression: A statistical mechanics analysis. *Advances in Neural Information Processing Systems*, 34.

[Mezard and Montanari, 2009] Mezard, M. and Montanari, A. (2009). *Information, physics, and computation*. Oxford University Press.

[Morcos et al., 2011] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301.

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

[Nguyen and Berg, 2012] Nguyen, H. C. and Berg, J. (2012). Bethe–Peierls approximation and the inverse Ising problem. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(03):P03004.

[Nguyen et al., 2017] Nguyen, H. C., Zecchina, R., and Berg, J. (2017). Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics*, 66(3):197–261.

[Nishimori, 2001] Nishimori, H. (2001). *Statistical physics of spin glasses and information processing: an introduction*. Number 111. Clarendon Press.

[Opper and Saad, 2001] Opper, M. and Saad, D. (2001). *Advanced mean field methods: Theory and practice*. MIT press.

[Plan and Vershynin, 2016] Plan, Y. and Vershynin, R. (2016). The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537.

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

[Ricci-Tersenghi, 2012] Ricci-Tersenghi, F. (2012). The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08015.

[Rockafellar, 1970] Rockafellar, R. T. (1970). *Convex analysis*, volume 36. Princeton university press.

[Rothman et al., 2008] Rothman, A. J., Bickel, P. J., Levina, E., Zhu, J., et al. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

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

[Scikit-learn,] Scikit-learn. Ridge classification. `https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression`.

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

[Thrampoulidis et al., 2015] Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2015). Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, 28:3420–3428.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Vershynin, 2018] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

[Vuffray et al., 2016] Vuffray, M., Misra, S., Lokhov, A., and Chertkov, M. (2016). Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603.

[Wainwright, 2009] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

[Wainwright et al., 2007] Wainwright, M. J., Lafferty, J. D., and Ravikumar, P. K. (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in neural information processing systems*, pages 1465–1472.

[Zhang et al., 2016] Zhang, Y., Guo, W., and Ray, S. (2016). On the consistency of feature selection with lasso for non-linear targets. In *International Conference on Machine Learning*, pages 183–191. PMLR.

[Zhao and Yu, 2006] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

A Proof of Lemma 1

Proof. The gradient of the square loss $\ell\left(\theta_{\backslash r}; \mathfrak{X}_1^n\right)$ in (5) w.r.t. $\theta_{\backslash r}$ reads

$$\nabla \ell\left(\theta_{\backslash r}; \mathfrak{X}_1^n\right) = \frac{1}{n}\sum_{i=1}^{n} x_{\backslash r}^{(i)}\left(x_r^{(i)} - \sum_{t\in V\backslash r}\theta_{rt}x_t^{(i)}\right). \tag{32}$$

After taking expectation of gradient $\nabla \ell\left(\theta_{\backslash r}; \mathfrak{X}_1^n\right)$ over the distribution $\mathbb{P}_{\theta^*}(x)$ and setting it to be zero, we obtain $\mathbb{E}_{\theta^*}\left(\nabla \ell\left(\theta_{\backslash r}; \mathfrak{X}_1^n\right)\right) = 0$ in matrix form:

$$Q_r^* \theta_{\backslash r} = b, \tag{33}$$

where $Q_r^* = \mathbb{E}_{\theta^*}\left(X_{\backslash r}X_r^T\right)$ is the covariance matrix of $X_{\backslash r}$ and $b = \mathbb{E}_{\theta^*}\left(X_{\backslash r}X_r\right)$. The solution to (33), denoted as $\tilde{\theta}_{\backslash r}^*$, can be analytically obtained as $\tilde{\theta}_{\backslash r}^* = \left(Q_r^*\right)^{-1} b$. Next, we construct the full covariance matrix $C = \mathbb{E}_{\theta^*}\left(XX^T\right)$ of all spins X as follows

$$C = \begin{bmatrix} 1 & b^T \\ b & Q_r^* \end{bmatrix}, \tag{34}$$

where X_r is indexed as the first variable in C without loss of generality. From the block matrix inversion lemma, the inverse covariance matrix can be computed as

$$C^{-1} = \begin{bmatrix} F_{11}^{-1} & -F_{11}^{-1}\left(\tilde{\theta}_{\backslash r}^*\right)^T \\ -\tilde{\theta}_{\backslash r}^* F_{11}^{-1} & F_{22}^{-1} \end{bmatrix}, \tag{35}$$

where

$$F_{11} = 1 - b^T\left(Q_r^*\right)^{-1} b, \tag{36}$$

$$F_{22} = Q_r^* - bb^T. \tag{37}$$

On the other hand, for general tree-like graphs in the paramagnetic phase, the inverse covariance matrix C^{-1} can be computed from the Hessian of the Gibbs free energy [Ricci-Tersenghi, 2012, Nguyen and Berg, 2012, Abbara et al., 2020]. Specifically, each element of the covariance matrix $C = \{C_{rt}\}_{r,t\in V}$ can be expressed as

$$C_{rt} = \mathbb{E}_{\theta^*}(x_r x_t) - \mathbb{E}_{\theta^*}(x_r)\mathbb{E}_{\theta^*}(x_t) = \frac{\partial^2 \log Z(\sigma)}{\partial\sigma_r \partial\sigma_t}, \tag{38}$$

where $Z(\sigma) = \sum_x \mathbb{P}_{\theta^*}(x)\, e^{\sum_{s\in V}\sigma_s x_s}$ with $\sigma = \{\sigma_s\}_{s\in V}$ and the assessment is carried out at $\sigma = 0$. In addition, for technical convenience we introduce the Gibbs free energy as

$$A(m) = \max_{\sigma}\left\{\sigma^T m - \log Z(\sigma)\right\}. \tag{39}$$

The definition of (39) indicates that following two relations hold:

$$\frac{\partial m_r}{\partial \sigma_t} = \frac{\partial^2 \log Z(\sigma)}{\partial\sigma_r\partial\sigma_t} = C_{rt}, \tag{40}$$

$$\frac{\partial \sigma_r}{\partial m_t} = \left[C^{-1}\right]_{rt} = \frac{\partial^2 A(m)}{\partial m_r \partial m_t}, \tag{41}$$

where the evaluations are performed at $\sigma = 0$ and $m = \arg\min_m A(m)$ ($= 0$ under the paramagnetic assumption). Consequently, the inverse covariance matrix of a tree-like graph $G \in \mathcal{G}_{p,d}$ can be computed as [Ricci-Tersenghi, 2012, Nguyen and Berg, 2012, Abbara et al., 2020]

$$\begin{aligned}\left[C^{-1}\right]_{rt} = &\left(\sum_{u\in\mathcal{N}(r)}\frac{1}{1-\tanh^2\left(\theta_{ru}^*\right)} - d_r + 1\right)\delta_{rt} \\ &- \frac{\tanh\left(\theta_{rt}^*\right)}{1-\tanh^2\left(\theta_{rt}^*\right)}\left(1-\delta_{rt}\right).\end{aligned} \tag{42}$$

The two representations of C^{-1} in (35) and (42) are equivalent so that the corresponding elements should equal to each other. Thus, the following identities hold

$$\begin{cases} F_{11}^{-1} = \sum_{u\in\mathcal{N}(r)} \frac{1}{1-\tanh^2(\theta^*_{ru})} - d_r + 1, \\ \tilde{\theta}^*_{\setminus r} F_{11}^{-1} = \frac{\tanh\left(\theta^*_{\setminus r}\right)}{1-\tanh^2\left(\theta^*_{\setminus r}\right)}, \end{cases} \tag{43}$$

where $\tanh(\cdot)$ is applied element-wise. From (43), we obtain (8), which is a rescaled version of the true interactions. In particular, for RR graphs with constant coupling $\theta^*_{rt} = \theta_0, \forall (r,t) \in E$ and $d_r = d$, substituting the results one can obtain

$$\tilde{\theta}^*_{rt} = \begin{cases} \frac{\tanh(\theta_0)}{1+(d-1)\tanh^2(\theta_0)} & \text{if } (r,t) \in E; \\ 0 & \text{otherwise.} \end{cases} \tag{44}$$

which completes the proof. □

B Proof of Lemma 2

The corresponding belief propagation (BP) equation on a RR graph can be written as follows [Mezard and Montanari, 2009]

$$m_{r\to t} = \tanh\left(\sum_{k\in\mathcal{N}(r)\setminus t} \tanh^{-1}\left(\tanh(\theta_0)\, m_{k\to r}\right)\right). \tag{45}$$

where $m_{r\to t}$ is the message from node r to node t. The spontaneous magnetization for the node $r \in V$ is assessed as

$$m_r = \tanh\left(\sum_{t\in\mathcal{N}(r)} \tanh^{-1}\left(\tanh(\theta_0)\, m_{t\to r}\right)\right). \tag{46}$$

Due to the uniformity of RR graphs, these equations are reduced to

$$m_c = \tanh\left((d-1)\tanh^{-1}\left(\tanh(\theta_0)\, m_c\right)\right), \tag{47}$$

$$m = \tanh\left(d\tanh^{-1}\left(\tanh(\theta_0)\, m_c\right)\right), \tag{48}$$

where we set $m_{r\to t} \coloneqq m_c$ and $m_r \coloneqq m$ for all directed edges $r \to t$ and all nodes $r \in V$.

Suppose that $x = (x_r)_{r=1}^p$ is subject to a Hamiltonian $H(x) = -\sum_{s\neq t} \theta^*_{rt} x_r x_t$. For this, we define the Helmholtz free energy as

$$F(\xi) = -\ln\left(\sum_x \exp\left(-H(x) + \sum_{r=1}^p \xi_r x_r\right)\right). \tag{49}$$

Using $F(\xi)$, one can evaluate the expectation as

$$m_r \coloneqq \mathbb{E}_{\theta^*}\{x_r\} = -\left.\frac{\partial F(\xi)}{\partial \xi_r}\right|_{\xi=0} = \frac{\sum_x x_r \exp(-H(x))}{\sum_x \exp(-H(x))}. \tag{50}$$

In addition, the covariance of x_r and x_t can be computed as

$$\begin{aligned} &\mathbb{E}_{\theta^*}\{x_r x_t\} - \mathbb{E}_{\theta^*}\{x_r\}\mathbb{E}_{\theta^*}\{x_t\} = \left.\frac{\partial \mathbb{E}_{\theta^*}\{x_r\}}{\partial \xi_t}\right|_{\xi=0} \\ =&\frac{\sum_x x_r x_t \exp(-H(x))}{\sum_x \exp(-H(x))} - \frac{\sum_x x_r \exp(-H(x))}{\sum_x \exp(-H(x))} \cdot \frac{\sum_x x_t \exp(-H(x))}{\sum_x \exp(-H(x))}, \end{aligned} \tag{51}$$

where the last equation is termed the *linear response relation [Nishimori, 2001].*

Suppose that node r is placed at the distance of l from node t. A remarkable property of tree-like graphs, including typical RR graphs, is that a unique path is defined between two arbitrary nodes. This indicates that the linear response relation (51) can be evaluated by the chain rule of partial derivative using messages of belief propagation as

$$\begin{aligned}\mathbb{E}_{\theta^*}\{x_r x_t\} - \mathbb{E}_{\theta^*}\{x_r\}\mathbb{E}_{\theta^*}\{x_t\} &= \left.\frac{\partial m_r}{\partial \xi_t}\right|_{\xi=0} \\ &= (1-m^2)\left(\frac{\tanh(\theta_0)(1-m_c^2)}{1-\tanh^2(\theta_0)\, m_c^2}\right)^l. \end{aligned} \tag{52}$$

In the the paramagnetic phase where $m = 0$ and $m_c = 0$, we have

$$\mathbb{E}_{\theta^*}\{x_r x_t\} - \mathbb{E}_{\theta^*}\{x_r\}\mathbb{E}_{\theta^*}\{x_t\} = \tanh^l(\theta_0). \tag{53}$$

Let us examine the *dependency condition* (C1). Since the distances between any two different nodes in $S \coloneqq \{(r,t) \mid t \in \mathcal{N}(r)\}$ are 2, all the off-diagonal elements in sub-matrix Q^*_{SS} equal to $\tanh^2\theta_0$ and all the diagonal elements equal to 1, i.e.,

$$Q^*_{SS} = \begin{bmatrix} 1 & \tanh^2\theta_0 & \tanh^2\theta_0 & \cdots & \tanh^2\theta_0 \\ \tanh^2\theta_0 & 1 & \tanh^2\theta_0 & \vdots & \tanh^2\theta_0 \\ \tanh^2\theta_0 & \tanh^2\theta_0 & \ddots & \tanh^2\theta_0 & \vdots \\ \vdots & \cdots & \tanh^2\theta_0 & 1 & \tanh^2\theta_0 \\ \tanh^2\theta_0 & \tanh^2\theta_0 & \cdots & \tanh^2\theta_0 & 1 \end{bmatrix}_{d\times d}. \tag{54}$$

It can be analytically computed that Q^*_{SS} has two different eigenvalues: one is $1 + (d-1)\tanh^2\theta_0$ and the other is $1 - \tanh^2\theta_0$ with multiplicity $(d-1)$. Consequently, Q^*_{SS} has bounded eigenvalue and we explicitly obtain the result of $C_{\min}$ as

$$\Lambda_{\min}(Q^*_{SS}) = 1 - \tanh^2\theta_0 \coloneqq C_{\min}. \tag{55}$$

Then, we prove that the *incoherence condition* (C2) also satisfies. From (54), the inverse matrix $(Q^*_{SS})^{-1}$ can be analytically computed as

$$(Q^*_{SS})^{-1} = \begin{bmatrix} a & b & b & \cdots & b \\ b & a & b & \vdots & b \\ b & b & \ddots & b & \vdots \\ \vdots & \cdots & b & a & b \\ b & b & \cdots & b & a \end{bmatrix}_{d\times d}, \tag{56}$$

where

$$a = \frac{1 + (d-2)\tanh^2\theta_0}{\left(1-\tanh^2\theta_0\right)\left(1+(d-1)\tanh^2\theta_0\right)}, \tag{57}$$

$$b = -\frac{\tanh^2\theta_0}{\left(1-\tanh^2\theta_0\right)\left(1+(d-1)\tanh^2\theta_0\right)}. \tag{58}$$

Then, by definition of $|||Q^*_{S^cS}(Q^*_{SS})^{-1}|||_\infty$, it is achieved for $r \in S^c$ where r belongs to the nearest neighbors of the nodes in S. Specifically, in that case, the elements in the row in $Q^*_{S^cS}$ associated with node $r \in S^c$ can only take two different values: one element is $\tanh\theta_0$ and the other $(d-1)$ elements are $\tanh^3\theta_0$. Then, from (56), after some algebra, it can be calculated that

$$|||Q^*_{S^cS}(Q^*_{SS})^{-1}|||_\infty = \tanh\theta_0 \coloneqq 1 - \alpha, \tag{59}$$

where we obtain an analytical result $\alpha \coloneqq 1 - \tanh\theta_0 \in (0,1]$, which completes the proof.

C Proofs of the key results

C.1 Proof of Lemma 3

Proof. The result that $\mathbb{E}_{\theta^*}\left(Z_s^{(i)}\right) = 0$ can be readily obtained by the definition of $\tilde{\theta}^*_{\backslash r}$ in Lemma 1. Thus, to prove $\texttt{Var}\left(Z_s^{(i)}\right) \leq 1$, it suffices to prove $\mathbb{E}_{\theta^*}\left(\left(Z_s^{(i)}\right)^2\right) \leq 1$ in the paramagnetic phase.

We introduce an auxiliary function

$$f_1\left(\theta_{\backslash r}\right) = \mathbb{E}_{\theta^*}\left(x_r^{(i)} - \sum_{t\in V\backslash r}\theta_t x_t^{(i)}\right)^2. \tag{60}$$

Thus we have $\mathbb{E}_{\theta^*}\left(\left(Z_s^{(i)}\right)^2\right) = f_1\left(\tilde{\theta}^*_{\backslash r}\right)$. The gradient vector can be computed as $\nabla f_1\left(\theta_{\backslash r}\right) = 2\mathbb{E}_{\theta^*}\left(\nabla \ell\left(\theta_{\backslash r}; \mathfrak{X}_n\right)\right)$. Since $\mathbb{E}_{\theta^*}\left(\nabla \ell\left(\tilde{\theta}^*_{\backslash r}; \mathfrak{X}_n\right)\right) = 0$ as shown in Lemma 1, we have $\nabla f_1\left(\tilde{\theta}^*_{\backslash r}\right) = 0$. Moreover, since $\nabla^2 f_1\left(\theta_{\backslash r}\right) = 2\mathbb{E}_{\theta^*}\left(X_{\backslash r}X_{\backslash r}^T\right) \succ 0$, we can conclude that $f_1\left(\theta_{\backslash r}\right)$ reaches its minimum at $\theta_{\backslash r} = \tilde{\theta}^*_{\backslash r}$. As a result, we have

$$\begin{aligned}
\mathbb{E}_{\theta^*}\left(\left(Z_s^{(i)}\right)^2\right) =& f_1\left(\theta_{\backslash r} = \tilde{\theta}^*_{\backslash r}\right)\\
\leq& f_1\left(\theta_{\backslash r} = 0\right)\\
=& \mathbb{E}_{\theta^*}\left(x_r^{(i)}\right)^2\\
=& 1,
\end{aligned} \tag{61}$$

where in the last line the fact that $x_r^{(i)} \in \{-1, +1\}, \forall r \in V$ is used. Therefore, we obtain $\texttt{Var}\left(Z_s^{(i)}\right) \leq 1$.

Moreover, the absolute value $\left|Z_s^{(i)}\right|$ is bounded. Specifically, (a) for RR graphs, in the paramagnetic phase, we have

$$\begin{aligned}
\left|Z_s^{(i)}\right| &= \left|x_s^{(i)}(x_r^{(i)} - \sum_{t\in V\backslash r}\tilde{\theta}^*_{rt}x_t^{(i)})\right|\\
&\leq 1 + \sum_{t\in V\backslash r}|\tilde{\theta}^*_{rt}|\\
&= 1 + \frac{d\tanh\left(\theta_0\right)}{1 + (d-1)\tanh^2\left(\theta_0\right)}\\
&\leq 2.
\end{aligned} \tag{62}$$

(b) for general tee-like graphs, recalling the result (8), we have

$$\begin{aligned}
&\left(\sum_{u\in\mathcal{N}(r)}\frac{1}{1-\tanh^2\left(\theta^*_{ru}\right)} - d_r + 1\right)\sum_{t\in V\backslash r}\left|\tilde{\theta}^*_{rt}\right|\\
=& \sum_{t\in\mathcal{N}(r)}\frac{\left|\tanh\left(\theta^*_{rt}\right)\right|}{1-\tanh^2\left(\theta^*_{rt}\right)}\\
=& \sum_{t\in\mathcal{N}(r)}\frac{\left|\tanh\left(\theta^*_{rt}\right)\right| + 1 - \tanh^2\left(\theta^*_{rt}\right) + \tanh^2\left(\theta^*_{rt}\right) - 1}{1-\tanh^2\left(\theta^*_{rt}\right)}\\
=& -d_r + \sum_{t\in\mathcal{N}(r)}\frac{\left|\tanh\left(\theta^*_{rt}\right)\right| + 1 - \tanh^2\left(\theta^*_{rt}\right)}{1-\tanh^2\left(\theta^*_{rt}\right)},
\end{aligned} \tag{63}$$

To proceed, consider an auxiliary function $f_2(x) = x + 1 - x^2, 0 \leq x \leq 1$. Then it can be proved that $1 \leq f_2(x) \leq \frac{5}{4}$, so that from (63), we have

$$\sum_{t \in V \setminus r} \left| \tilde{\theta}^*_{rt} \right| \leq \frac{-d_r + \frac{5}{4} \sum_{u \in \mathcal{N}(r)} \frac{1}{1-\tanh^2(\theta^*_{ru})}}{\sum_{u \in \mathcal{N}(r)} \frac{1}{1-\tanh^2(\theta^*_{ru})} - d_r + 1}. \tag{64}$$

It can be easily checked that $\sum_{u \in \mathcal{N}(r)} \frac{1}{1-\tanh^2(\theta^*_{ru})} \in [d_r, \infty)$. We introduce another auxiliary function

$$f_3(x) = \frac{-d_r + \frac{5}{4}x}{x - d_r + 1}, x \in [d_r, \infty). \tag{65}$$

The first-order derivative of $f_3(x)$ can be easily computed as

$$f_3^{'}(x) = \frac{5 - d_r}{4(x - d_r + 1)^2}. \tag{66}$$

As a result, $f_3^{'}(x) > 0$ when $d_r < 5$ and $f_3^{'}(x) < 0$ when $d_r > 5$. Consequently,

$$\max_{x \in [d_r, \infty)} f_3(x) = \begin{cases} \frac{5}{4} & d_r \leq 5 \\ \frac{d_r}{4} & d_r > 5 \end{cases} \tag{67}$$

Finally, combining the above results together yields

$$\left| Z_s^{(i)} \right| \leq \max \left\{ \frac{9}{4}, \frac{4 + d_r}{4} \right\} < d_r, \forall d_r \geq 3. \tag{68}$$

By definition, there is $d_r \leq d$ so that $\left| Z_s^{(i)} \right| \leq d$, which completes the proof. □

C.2 Proof of Lemma 4

Proof. Frist, we prove the case (a). In this case, According to Lemma 3, $\mathbb{E}_{\theta^*}\left(Z_s^{(i)}\right) = 0$ and $\left| Z_s^{(i)} \right| \leq 2$, so that by the Azuma Hoeffding inequality [Vershynin, 2018], for $\forall \eta > 0$, we have

$$\mathbb{P}\left(|W_s^n| > \eta\right) \leq 2\exp\left(-\frac{\eta^2 n}{8}\right). \tag{69}$$

Setting $\eta = \frac{\alpha \lambda_n}{2(2-\alpha)}$, we obtain

$$\mathbb{P}\left(\frac{2-\alpha}{\lambda_n}|W_s^n| > \frac{\alpha}{2}\right) \leq 2\exp\left(-\frac{\alpha^2 \lambda_n^2 n}{32(2-\alpha)^2}\right). \tag{70}$$

Then, by using a union bound we have

$$\mathbb{P}\left(\frac{2-\alpha}{\lambda_n}\|W^n\|_\infty \geq \frac{\alpha}{2}\right) \leq 2\exp\left(-\frac{\alpha^2 \lambda_n^2 n}{32(2-\alpha)^2} + \log p\right), \tag{71}$$

which completes the proof of (a).

In the case (b) for general graphs, the proof is slightly complicated. According to Lemma 3, applying the Bernstein's inequality [Vershynin, 2018], $\forall \eta > 0$ we have

$$\mathbb{P}\left(|W_s^n| > \eta\right) \leq 2\exp\left(-\frac{\frac{1}{2}\eta^2 n}{1 + \frac{1}{3}d\eta}\right). \tag{72}$$

Similar to [Vuffray et al., 2016], inverting the following relation

$$\xi = \frac{\frac{1}{2}\eta^2 n}{1 + \frac{1}{3}d\eta}, \tag{73}$$

and substituting the result in (72) yields

$$\mathbb{P}\left(|W_s^n| > \frac{1}{3}\left(u + \sqrt{u^2 + 18\frac{u}{d}}\right)\right) \leq 2\exp\left(-\xi\right), \tag{74}$$

where $u = \frac{\xi}{n}d$. Suppose that $n \geq \xi d^2$, then $u^2 = \frac{\xi^2}{n^2}d^2 \leq \frac{\xi}{n}$ while $\frac{u}{d} = \frac{\xi}{n}$. Consequently, we have

$$\begin{aligned}\frac{1}{3}\left(u + \sqrt{u^2 + 18\frac{u}{d}}\right) &\leq \frac{1}{3}\left(\sqrt{\frac{\xi}{n}} + \sqrt{\frac{\xi}{n} + 18\frac{\xi}{n}}\right) &(75)\\ &\leq \frac{1}{3}\left(\sqrt{\frac{\xi}{n}} + \sqrt{\frac{\xi}{n}}\sqrt{25}\right) &(76)\\ &= 2\sqrt{\frac{\xi}{n}}, &(77)\end{aligned}$$

where a relaxed result is obtained. Subsequently, we obtain an expression which is independent of d:

$$\mathbb{P}\left(|W_s^n| > 2\sqrt{\frac{\xi}{n}}\right) \leq 2\exp\left(-\xi\right). \tag{78}$$

Setting $\xi = (c+1)\log p$, then if $\lambda_n \geq \frac{4(2-\alpha)\sqrt{c+1}}{\alpha}\sqrt{\frac{\log p}{n}}$, we have $\frac{\alpha\lambda_n}{2(2-\alpha)} \geq 2\sqrt{\frac{\xi}{n}}$ so that

$$\begin{aligned}&\mathbb{P}\left(\frac{2-\alpha}{\lambda_n}|W_s^n| > \frac{\alpha}{2}\right) \leq \mathbb{P}\left(|W_s^n| > 2\sqrt{\frac{\xi}{n}}\right)\\ &\leq 2\exp\left(-\left(c+1\right)\log p\right).\end{aligned} \tag{79}$$

Then, by using a union bound we have

$$\mathbb{P}\left(\frac{2-\alpha}{\lambda_n}\left\|W^n\right\|_\infty \geq \frac{\alpha}{2}\right) \leq 2\exp\left(-c\log p\right). \tag{80}$$

As a result, when $n \geq (c+1)\,d^2\log p$, as long as $\lambda_n \geq \frac{4\sqrt{c+1}(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$, it is guaranteed that $\mathbb{P}\left(\frac{2-\alpha}{\lambda_n}\left\|W^n\right\|_\infty \geq \frac{\alpha}{2}\right) \to 0$ at rate $\exp\left(-c\log p\right)$ for some constant $c > 0$, which completes the proof. □

C.3 Proof of Lemma 5

Proof. Using the method in [Rothman et al., 2008], here the proof follows [Ravikumar et al., 2010] but with essential modifications. First, define a function $\mathbb{R}^d \to \mathbb{R}$ as follows [Rothman et al., 2008]

$$\begin{aligned}G\left(u_S\right) \coloneqq &\ell\left(\tilde{\theta}_S^* + u_S; \mathfrak{X}_n\right) - \ell\left(\tilde{\theta}_S^*; \mathfrak{X}_n\right)\\ &+ \lambda_n\left(\left\|\tilde{\theta}_S^* + u_S\right\|_1 - \left\|\tilde{\theta}_S^*\right\|_1\right).\end{aligned} \tag{81}$$

Note that G is a convex function w.r.t. u_S. Then $\hat{u}_S = \hat{\theta}_S - \tilde{\theta}_S^*$ minimizes G according to the definition in (4). Moreover, it is easily seen that $G\left(0\right) = 0$ so that $G\left(\hat{u}_S\right) \leq 0$. As described in [Ravikumar et al., 2010], if we can show that there exists some radius $B > 0$ and any $u_S \in \mathbb{R}^d$ with $\left\|u_S\right\|_2 = B$ satisfies $G(u_S) > 0$, then we can claim that $\left\|\hat{u}_S\right\|_2 \leq B$ since otherwise one can always, by appropriately choosing $t \in (0,1]$, find a convex combination $t\hat{u}_S + (1-t)\,0$ which lies on the boundary of the ball with radius B and thus $G\left(t\hat{u}_S + (1-t)\,0\right) \leq 0$, leading to contradiction. Consequently, it suffices to establish the strict positivity of G on the boundary of a ball with radius $B = M\lambda_n\sqrt{d}$, where $M > 0$ is one parameter to choose later.

Specifically, let $u_S \in \mathbb{R}^d$ be an arbitrary vector with $\left\|u_S\right\|_2 = B$. Expanding the quadratic form $\ell\left(\tilde{\theta}_S^* + u_S; \mathfrak{X}_n\right)$, we have

$$\begin{aligned}G\left(u_S\right) = &-\left(W_S^n\right)^T u_S + u_S^T Q_{SS}^n u_S\\ &+ \lambda_n\left(\left\|\tilde{\theta}_S^* + u_S\right\|_1 - \left\|\tilde{\theta}_S^*\right\|_1\right),\end{aligned} \tag{82}$$

where W_S^n is the sub-vector of $W^n = -\nabla \ell\left(\tilde{\theta}^*; \mathfrak{X}_n\right)$, and Q_{SS}^n is the sub-matrix of the sample matrix Q^n. The expression (82) is simpler than the counterpart in [Ravikumar et al., 2010] which is obtained from the Taylor series expansion of the non-quadratic loss function and thus its quadratic term is dependent on θ. To proceed, we investigate the bounds of the three terms in the right hand side (RHS) of (82), respectively.

Since $\|u_S\|_1 \leq \sqrt{d}\,\|u_S\|_2$ and $\|W_S^n\|_\infty \leq \frac{\lambda_n}{2}$, the first term is bounded as

$$\begin{aligned}\left|-\left(W_S^n\right)^T u_S\right| &\leq \|W_S^n\|_\infty \|u_S\|_1 \leq \|W_S^n\|_\infty \sqrt{d}\,\|u_S\|_2 \\ &\leq \left(\lambda_n\sqrt{d}\right)^2 \frac{M}{2}. \end{aligned} \tag{83}$$

The third term is bounded as

$$\begin{aligned}&\lambda_n\left(\left\|\tilde{\theta}_S^* + u_S\right\|_1 - \left\|\tilde{\theta}_S^*\right\|_1\right) \\ &\geq -\lambda_n \|u_S\|_1 \geq -\lambda_n\sqrt{d}\,\|u_S\|_2 \\ &= -M\left(\lambda_n\sqrt{d}\right)^2. \end{aligned} \tag{84}$$

The remaining middle Hessian term in RHS of (82) is, different from [Ravikumar et al., 2010], quite simple due to the square loss function:

$$\begin{aligned}u_S^T Q_S^n u_S &\geq \|u_S\|_2^2\,\Lambda_{\min}\left(Q_{SS}^n\right) \\ &\geq C_{\min} M^2\left(\lambda_n\sqrt{d}\right)^2, \end{aligned} \tag{85}$$

where the last inequality comes from the dependency condition $\Lambda_{\min}\left(Q_{SS}^*\right) \geq C_{\min}$ in (13). In contrast to [Ravikumar et al., 2010], there is no need to control the additional spectral norm.

Combining the three bounds (83) - (85) together with (82), we obtain that

$$G\left(u_S\right) \geq \left(\lambda_n\sqrt{d}\right)^2\left\{-\frac{M}{2} + C_{\min}M^2 - M\right\}. \tag{86}$$

It can be easily verified from (86) that $G\left(u_S\right)$ is strictly positive when we choose $M = \frac{3}{C_{\min}}$. Consequently, as long as $\|W^n\|_\infty \leq \frac{\lambda_n}{2}$, we are guaranteed that $\|\hat{u}_S\|_2 \leq M\lambda_n\sqrt{d} = \frac{3\lambda_n\sqrt{d}}{C_{\min}}$, which completes the proof. □

C.4 Proof of Lemma 6

Proof. According to Lemma 3, applying the Bernstein's inequality, $\forall \eta > 0$ we have

$$\mathbb{P}\left(|W_s^n| > \eta\right) \leq 2\exp\left(-\frac{\frac{1}{2}\eta^2 n}{1 + \frac{1}{3}d\eta}\right). \tag{87}$$

Similar to [Vuffray et al., 2016], inverting the following relation

$$\xi = \frac{\frac{1}{2}\eta^2 n}{1 + \frac{1}{3}d\eta} \tag{88}$$

and substituting the result in (87) yields

$$\mathbb{P}\left(|W_s^n| > \frac{1}{3}\left(u + \sqrt{u^2 + 18\frac{u}{d}}\right)\right) \leq 2\exp\left(-\xi\right). \tag{89}$$

where $u = \frac{\xi}{n} d$. Suppose that $n \geq \xi d^2$, then $u^2 = \frac{\xi^2}{n^2} d^2 \leq \frac{\xi}{n}$ while $\frac{u}{d} = \frac{\xi}{n}$. Consequently, we have

$$\frac{1}{3}\left(u + \sqrt{u^2 + 18\frac{u}{d}}\right) \leq \frac{1}{3}\left(\sqrt{\frac{\xi}{n}} + \sqrt{\frac{\xi}{n} + 18\frac{\xi}{n}}\right) \tag{90}$$

$$\leq \frac{1}{3}\left(\sqrt{\frac{\xi}{n}} + \sqrt{\frac{\xi}{n}}\sqrt{25}\right) \tag{91}$$

$$= 2\sqrt{\frac{\xi}{n}}. \tag{92}$$

where a relaxed result is obtained. Subsequently, we obtain an expression which is independent of d

$$\mathbb{P}\left(|W_s^n| > 2\sqrt{\frac{\xi}{n}}\right) \leq 2\exp\left(-\xi\right). \tag{93}$$

Then, by using a union bound we have

$$\mathbb{P}\left(\|W^n\|_\infty > 2\sqrt{\frac{\xi}{n}}\right) \leq 2\exp\left(-\xi + \log p\right). \tag{94}$$

Setting $\xi = \log\frac{2p}{\varepsilon_3}$, then if $n \geq d^2 \log\frac{2p}{\varepsilon_3}$, we have

$$\mathbb{P}\left(\|W^n\|_\infty > 2\sqrt{\frac{\log\frac{2p}{\varepsilon_3}}{n}}\right) \leq 2\exp\left(-\log\frac{2p}{\varepsilon_3} + \log p\right) \tag{95}$$

$$= \varepsilon_3, \tag{96}$$

which completes the proof. □

C.5 Proof of Lemma 7

Proof. Since $x_r^{(i)} x_t^{(i)}$ is bounded by $\left|x_r^{(i)} x_t^{(i)}\right| \leq 1$. Therefore, using the Hoeffding inequality [Hoeffding, 1994], for any $\epsilon > 0$, there is

$$\mathbb{P}\left(|Q_{st}^n - Q_{st}^*| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right). \tag{97}$$

Then, due to the symmetry of Q_{st}^n, using a union bound we have

$$\mathbb{P}\left(|Q_{st}^n - Q_{st}^*| \leq \epsilon, \forall s,t \in V \setminus r\right) \geq 1 - p^2\exp\left(-\frac{n\epsilon^2}{2}\right), \tag{98}$$

As a result, as long as $n \geq \frac{2}{\epsilon^2}\log\frac{p^2}{\varepsilon_4}$, there is $\mathbb{P}\left(|Q_{st}^n - Q_{st}^*| \leq \epsilon, \forall s,t \in V \setminus r\right) \geq 1 - \varepsilon_4$, which completes the proof. □

C.6 Proof of Lemma 9

Proof. According (28) and Lemma 8, we have

$$\begin{aligned}
&\delta\ell\left(\triangle_{\theta_{\setminus r}}, \tilde{\theta}^*_{\setminus r}; \mathfrak{X}_1^n\right) \\
&= \frac{1}{2}\triangle_{\theta_{\setminus r}}^T Q^n \triangle_{\theta_{\setminus r}} \\
&= \frac{1}{2}\triangle_{\theta_{\setminus r}}^T Q^* \triangle_{\theta_{\setminus r}} + \frac{1}{2}\triangle_{\theta_{\setminus r}}^T \left(Q^n - Q^*\right)\triangle_{\theta_{\setminus r}} \\
&\geq \frac{e^{-2\theta^*_{\max}d}}{2\left(d+1\right)}\left\|\triangle_{\theta_{\setminus r}}\right\|_2^2 + \frac{1}{2}\triangle_{\theta_{\setminus r}}^T \left(Q^n - Q^*\right)\triangle_{\theta_{\setminus r}}.
\end{aligned} \tag{99}$$

Then, from Lemma 7, choosing $\epsilon = \frac{e^{-2\theta^*_{\max}d}}{32d(d+1)}$, then with probability at least $1-\varepsilon_4$, there is

$$\begin{aligned}\triangle^T_{\theta_{\backslash r}}\left(Q^n - Q^*\right)\triangle_{\theta_{\backslash r}} &\geq -\frac{e^{-2\theta^*_{\max}d}}{32d\left(d+1\right)}\left\|\triangle_{\theta_{\backslash r}}\right\|_1^2 \\ &\geq -\frac{e^{-2\theta^*_{\max}d}}{2\left(d+1\right)}\left\|\triangle_{\theta_{\backslash r}}\right\|_2^2. \end{aligned} \tag{100}$$

as long as $n \geq \frac{2}{\epsilon^2}\log\frac{p^2}{\varepsilon_4} = 2^{11}d^2\left(d+1\right)^2 e^{4\theta^*_{\max}d}\log\frac{p^2}{\varepsilon_4}$. As a result, there is

$$\begin{aligned}&\delta\ell\left(\triangle_{\theta_{\backslash r}}, \tilde{\theta}^*_{\backslash r}; \mathfrak{X}_1^n\right) \\ &\geq \frac{e^{-2\theta^*_{\max}d}}{2\left(d+1\right)}\left\|\triangle_{\theta_{\backslash r}}\right\|_2^2 - \frac{e^{-2\theta^*_{\max}d}}{4\left(d+1\right)}\left\|\triangle_{\theta_{\backslash r}}\right\|_2^2 \\ &= \frac{e^{-2\theta^*_{\max}d}}{4\left(d+1\right)}\left\|\triangle_{\theta_{\backslash r}}\right\|_2^2, \end{aligned} \tag{101}$$

which completes the proof. □

D Proofs of Theorems 1 and 2

First, to prove the "fixed design" results in Proposition 1 and Proposition 2, for each vertex $r \in V$, an optimal primal-dual pair $\left(\hat{\theta}_{\backslash r}, \hat{z}_r\right)$ is constructed, where $\hat{\theta}_{\backslash r} \in \mathbb{R}^{p-1}$ is a primal solution and $\hat{z}_r \in \mathbb{R}^{p-1}$ is the associated sub-gradient vector. They satisfy the zero sub-gradient optimality condition [Rockafellar, 1970] associated with Lasso (4):

$$\nabla\ell\left(\hat{\theta}_{\backslash r}; \mathfrak{X}_n\right) + \lambda_n\hat{z}_r = 0, \tag{102}$$

where the sub-gradient vector $\hat{z}_r$ satisfies

$$\begin{cases}\hat{z}_{rt} = \text{sign}\left(\hat{\theta}_{rt}\right), \text{if } \hat{\theta}_{rt} \neq 0; & (a)\\ |\hat{z}_{rt}| \leq 1, \text{otherwise.} & (b)\end{cases} \tag{103}$$

Then, the pair is a primal-dual optimal solution to (4) and its dual. Further, to ensure that such an optimal primal-dual pair correctly specifies the signed neighborbood of node r, the sufficient and necessary conditions are as follows

$$\begin{cases}\text{sign}\left(\hat{z}_{rt}\right) = \text{sign}\left(\theta^*_{rt}\right), \forall\left(r,t\right) \in S, & (a)\\ \hat{\theta}_{ru} = 0, \forall\left(r,u\right) \in S^c \coloneqq E \setminus S. & (b)\end{cases} \tag{104}$$

Note that while the regression in (4) corresponds to a convex problem, for $p \gg n$ in the high-dimensional regime, it is not necessarily strictly convex so that there might be multiple optimal solutions. Fortunately, the following lemma in [Ravikumar et al., 2010] provides sufficient conditions for shared sparsity among optimal solutions as well as uniqueness of the optimal solution.

Lemma 10. *(Lemma 1 in [Ravikumar et al., 2010]). Suppose that there exists an optimal primal solution $\hat{\theta}_{\backslash r}$ with associated optimal dual vector $\hat{z}_r$ such that $\|\hat{z}_{S^c}\|_\infty < 1$. Then any optimal primal solution $\tilde{\theta}$ must have $\tilde{\theta}_{S^c} = 0$. Moreover, if the Hessian sub-matrix $[\nabla^2\ell\left(\hat{\theta}_{\backslash r}; \mathfrak{X}_n\right)]_{SS}$ is strictly positive definite, then $\hat{\theta}_{\backslash r}$ is the unique optimal solution.*

As a result, using the framework in [Ravikumar et al., 2010], we can construct a primal-dual witness $\left(\hat{\theta}_{\backslash r}, \hat{z}\right)$ for the Lasso estimator (4) as follows:

(a) First, set $\hat{\theta}_S$ as the minimizer of the partial penalized likelihood

$$\hat{\theta}_S = \underset{\theta_{\backslash r}=(\theta_S,0)\in\mathbb{R}^{p-1}}{\arg\min}\left\{\ell\left(\theta_{\backslash r}; \mathfrak{X}_n\right) + \lambda_n\left\|\theta_S\right\|_1\right\}, \tag{105}$$

and then set $\hat{z}_S = \text{sign}\left(\hat{\theta}_S\right)$.

(b) Second, set $\hat{\theta}_{S^c} = 0$ so that condition (104) (b) holds.

(c) Third, obtain $\hat{z}_{S^c}$ from (102) by substituting the values of $\hat{\theta}_{\backslash r}$ and $\hat{z}_S$.

(d) Finally, we need to show that the stated scalings of (n, p, d) imply that, with high probability, the remaining conditions (103) and (104) (a) are satisfied.

D.1 Proof of Proposition 1

From Lemma 4 (a), if the regularization parameter λ_n satisfies $\lambda_n \geq \frac{8(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$, then with probability greater than $1 - 2\exp\left(-c\lambda_n^2 n\right)$ there is

$$\|W^n\|_\infty \leq \frac{\alpha}{2-\alpha}\frac{\lambda_n}{2} \leq \frac{\lambda_n}{2}, \tag{106}$$

so that the condition in Lemma 5 is also satisfied. The zero-subgradient condition (102) can be equivalently re-written as follows

$$\begin{cases} Q^n_{S^cS}\left(\hat{\theta}_S - \tilde{\theta}^*_S\right) = W^n_{S^c} - \lambda_n \hat{z}_{S^c}, \\ Q^n_{SS}\left(\hat{\theta}_S - \tilde{\theta}^*_S\right) = W^n_S - \lambda_n \hat{z}_S, \end{cases} \tag{107}$$

where we have used the fact that $\hat{\theta}_{S^c} = 0$ from the primal-dual construction, and also the result $\tilde{\theta}^*_{S^c} = 0$ from Lemma 1. After some simple algebra, we obtain

$$W^n_{S^c} - Q^n_{S^cS}\left(Q^n_{SS}\right)^{-1} W^n_S + \lambda_n Q^n_{S^cS}\left(Q^n_{SS}\right)^{-1}\hat{z}_S = \lambda_n \hat{z}_{S^c}. \tag{108}$$

For strict dual feasibility, from (108), we obtain

$$\begin{aligned} \|\hat{z}_{S^C}\|_\infty &\leq |||Q^*_{S^CS}\left(Q^*_{SS}\right)^{-1}|||_\infty \left[\frac{\|W^n_S\|_\infty}{\lambda_n} + 1\right] \\ &\quad + \frac{\|W^n_{S^C}\|_\infty}{\lambda_n} \\ &\leq (1-\alpha) + (2-\alpha)\frac{\|W^n\|_\infty}{\lambda_n} \\ &\leq (1-\alpha) + (2-\alpha)\frac{1}{2-\alpha}\frac{\alpha}{2} \\ &= 1 - \frac{\alpha}{2} < 1, \end{aligned} \tag{109}$$

with probability converging to one. For correct sign recovery, it suffices to show that $\left\|\hat{\theta}_S - \tilde{\theta}^*_S\right\|_\infty \leq \frac{\tilde{\theta}^*_{\min}}{2}$. From Lemma 5 (since (106) holds), we have

$$\frac{2}{\theta^*_{\min}}\left\|\hat{\theta}_S - \tilde{\theta}^*_S\right\|_\infty \leq \frac{2}{\theta^*_{\min}}\left\|\hat{\theta}_S - \tilde{\theta}^*_S\right\|_2 \leq \frac{6}{\tilde{\theta}^*_{\min}C_{\min}}\lambda_n\sqrt{d}. \tag{110}$$

As a result, if $\tilde{\theta}^*_{\min} \geq \frac{6\lambda_n\sqrt{d}}{C_{\min}}$, or $\lambda_n \leq \frac{\tilde{\theta}^*_{\min}C_{\min}}{6\sqrt{d}}$, the condition $\left\|\hat{\theta}_S - \tilde{\theta}^*_S\right\|_\infty \leq \frac{\tilde{\theta}^*_{\min}}{2}$ holds. In the paramagnetic phase, from Lemma 1, there is $\tilde{\theta}^*_{\min} = \frac{\tanh(\theta_0)}{1+(d-1)\tanh^2(\theta_0)}$. Substituting these results lead to Proposition 1.

D.2 Proof of Proposition 2

The proof of Proposition 2 is the same as that of Proposition 1 in Appendix D.1, except that different conditions in Lemma 4 (b) are used, and that we need to impose the assumptions that the population Hessian Q^* satisfies both conditions (C1) and (C2) for the considered general graphs.

D.3 Proof of Theorem 1

Now we are ready to prove the main results in Theorem 1. As shown in Lemma 2, for RR graphs with uniform couplings, the population Hessian Q^* for Lasso already satisfies both conditions (C1) and (C2), so that assumptions of (C1) and (C2) can be dropped for RR graphs.

Next, using large deviation analysis as [Ravikumar et al., 2010], we prove that the sample Hessian Q^n of Lasso satisfies the same properties as the population Hessian Q^* with high probability with large enough samples.

Lemma 11. *Consider an Ising model on a RR graph $G = (V, E) \in \mathcal{G}_{p,d}$ with regular node degree d and uniform couplings $\theta^*_{r,t} = \theta_0, \forall (r,t) \in E$. Then, for any $\delta > 0$, there are some positive constants A, B, K*

$$\mathbb{P}\left(\Lambda_{\min}\left(Q^n_{SS}\right) \leq C_{\min} - \delta\right) \leq 2\exp\big(-A\frac{\delta^2 n}{d^2} + B\log d\big), \tag{111}$$

$$\mathbb{P}\left(\left|\left|\left|Q^n_{S^cS}\left(Q^n_{SS}\right)^{-1}\right|\right|\right|_\infty \geq 1 - \frac{\alpha}{2}\right) \leq 2\exp\big(-K\frac{n}{d^3} + \log p\big), \tag{112}$$

where $C_{\min}$ and α are $C_{\min} = 1 - \tanh^2\theta_0$ and $\alpha = 1 - \tanh\theta_0$.

Proof. The proof is the same as Lemma 5 and Lemma 6 in [Ravikumar et al., 2010], with the only difference that the variance function term does not exist, by substituting into the the results of $C_{\min}$ and α in the Lemma 2. □

Lemma 11 demonstrates that the sample Hessian Q^n satisfies both conditions (C1) and (C2) with high probability as long as $n \geq Ld^3 \log p$ for some constant L. As the results of Proposition 1 builds on top of the assumption that the sample Hessian Q^n satisfies (C1) and (C2), we readily obtain that all results of Proposition 1 will hold for if we replace the requirement that the sample Hessian Q^n satisfies both conditions (C1) and (C2) by an extra scaling requirement $n \geq Ld^3 \log p$ for some constant L independent of (n, p, d).

Consequently, by combining Lemma 2, Lemma 11, and Proposition 1 and substituting the specific results of $C_{\min}$ and α in Lemma 2, after some algebra, we readily obtain Theorem 1, which completes the proof.

D.4 Proof of Theorem 2

The proof of Theorem 2 is the same as that of Theorem 1 in Appendix D.3, except that different conditions in Lemma 4 (b) are used.

E Proofs of Theorems 3 and 4

E.1 Proofs of Theorem 3

This is done through Proposition 3 by evaluating the two conditions (C3) and (C4). First, let $\varepsilon_3 = \frac{2\varepsilon_1}{3} > 0$ in Lemma 6. Then, by setting $\lambda_n = 4\sqrt{\frac{\log\frac{3p}{\varepsilon_1}}{n}}$, if $n \geq d^2 \log\frac{3p}{\varepsilon_1}$, with probability at least $1 - \frac{2\varepsilon_1}{3}$, we have $\|W^n\|_\infty \leq 2\sqrt{\frac{\log\frac{3p}{\varepsilon_1}}{n}} = \frac{\lambda_n}{2}$ so that condition (C3) satisfies as long as $n \geq d^2 \log\frac{3p}{\varepsilon_1}$. Second, let $\varepsilon_4 = \frac{\varepsilon_1}{3} > 0$ in Lemma 9. From Lemma 9, with probability at least $1 - \frac{\varepsilon_1}{3}$, the restricted strong convexity condition is satisfied with the value $\kappa = \frac{e^{-2\theta^*_{\max}d}}{4(d+1)}$ when $n > 2^{11}d^2\left(d+1\right)^2 e^{4\theta^*_{\max}d}\log\frac{3p^2}{\varepsilon_1}$. Then, the relation $R \geq 3\sqrt{d}\frac{\lambda_n}{\kappa}$ in Proposition 3 reads

$$R > 3\sqrt{d}4\sqrt{\frac{\log\frac{3p}{\varepsilon_1}}{n}}\left(\frac{e^{-2\theta^*_{\max}d}}{4\left(d+1\right)}\right)^{-1}. \tag{113}$$

To find a value of R that satisfies (113), we can choose $R = 2/\sqrt{d}$. Then from (113), the number of samples n needs to satisfy

$$n > 9\cdot 2^{10}d^2\left(d+1\right)^2 e^{4\theta^*_{\max}d}\log\frac{3p^2}{\varepsilon_1}. \tag{114}$$

As a result, when $n \geq 2^{14}d^2\left(d+1\right)^2 e^{4\theta^*_{\max}d}\log\frac{3p^2}{\varepsilon_1}$, the condition (C4) satisfies with probability at least $1 - \frac{\varepsilon_1}{3}$. Based on the union bound, both condition (C3) and condition (C4) will be simultaneously satisfied with probability at least $1 - \varepsilon_1$, which completes the proof by using Proposition 3.

E.2 Proofs of Theorem 4

First consider any fixed vertex $r \in V$, if the square error $\left\|\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}\right\|_2 \leq \frac{\tilde{\theta}^*_{\min}}{2}$, then it is guaranteed that the absolute difference of each element of $\hat{\theta}_{\backslash r}$ and $\tilde{\theta}^*_{\backslash r}$ is less than $\frac{\tilde{\theta}^*_{\min}}{2}$ so that one can perfectly recover all its correct neighbors with a thresholding $\frac{\tilde{\theta}^*_{\min}}{2}$. According to Theorem 3, with probability $1-\varepsilon_1$, when $n \geq 2^{14} d^2 (d+1)^2 e^{4\theta^*_{\max} d} \log \frac{3p^2}{\varepsilon_1}$, there is $\left\|\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}\right\|_2 \leq 2^6 \sqrt{d}\,(d+1)\, e^{2\theta^*_{\max} d} \sqrt{\frac{\log \frac{3p}{\varepsilon_1}}{n}}$. Further, let $2^6 \sqrt{d}\,(d+1)\, e^{2\theta^*_{\max} d} \sqrt{\frac{\log \frac{3p}{\varepsilon_1}}{n}} \leq \frac{\tilde{\theta}^*_{\min}}{2}$, we obtain that $n \geq 2^{14} \left(\tilde{\theta}^*_{\min}\right)^{-2} d\,(d+1)^2\, e^{4\theta^*_{\max} d} \log \frac{3p}{\varepsilon_1}$. Consequently, with at least probability $1-\varepsilon_1$ we have $\left\|\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}\right\|_2 \leq \frac{\tilde{\theta}^*_{\min}}{2}$ and thus correct neighbors are recovered for any fixed $r \in V$ whenever

$$n \geq \max\left\{d, \left(\tilde{\theta}^*_{\min}\right)^{-2}\right\} 2^{14} d\,(d+1)^2\, e^{4\theta^*_{\max} d} \log \frac{3p^2}{\varepsilon_1}. \tag{115}$$

Then, setting $\varepsilon_2 = p\varepsilon_1$ and using the union bound for all vertices $r \in V$, we have

$$\mathbb{P}\left(\left\|\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}\right\|_2 > \frac{\tilde{\theta}^*_{\min}}{2}, \exists\, r \in V\right) \leq p\varepsilon_1 = \varepsilon_2, \tag{116}$$

so that

$$\mathbb{P}\left(\left\|\hat{\theta}_{\backslash r} - \tilde{\theta}^*_{\backslash r}\right\|_2 \leq \frac{\tilde{\theta}^*_{\min}}{2}, \forall\, r \in V\right) > 1 - \varepsilon_2, \tag{117}$$

which completes the proof.