Ising Model Selection Using ℓ_1-Regularized Linear Regression: A Statistical Mechanics Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

We theoretically investigate the typical learning performance of ℓ_1-regularized linear regression (ℓ_1-LinR) for Ising model selection using the replica method from statistical mechanics. For typical random regular (RR) graphs in the paramagnetic phase, we obtain an accurate estimate of the typical sample complexity of ℓ_1-LinR, which demonstrates that ℓ_1-LinR is model selection consistent with $M = \mathcal{O}(\log N)$ samples, where N is the number of variables of the Ising model. Moreover, we provide a computationally efficient method to accurately predict the non-asymptotic behavior of ℓ_1-LinR for moderate M and N, such as the precision and recall rates. Simulations show a fairly good agreement between the theoretical predictions and experimental results, even for graphs with many loops, which supports our findings. Although this paper focuses on ℓ_1-LinR, our method is readily applicable to precisely investigating the typical learning performance of a wide class of ℓ_1-regularized estimators for Ising model selection such as ℓ_1-regularized logistic regression and interaction screening.

1 Introduction

The advent of massive data across various scientific disciplines has led to the widespread use of undirected graphical models, also known as Markov random fields (MRFs), as a tool for discovering and visualizing dependencies among covariates in multivariate data [1]. The Ising model, originally proposed in statistical physics, is one special class of binary MRFs with pairwise potentials and has been widely used in different domains such as image analysis, social networking, gene network analysis [2, 3, 4, 5, 6, 7]. Among various applications, one fundamental problem of interest is called Ising model selection, which refers to recovering the underlying graph structure of the original Ising model from independent, identically distributed (i.i.d.) samples. A variety of methods have been proposed [8, 9, 10, 11, 12, 13, 14, 15, 16], demonstrating the possibility of successful Ising model selection even when the number of samples is smaller than that of variables. Notably, it has been demonstrated that for the ℓ_1-regularized logistic regression (ℓ_1-LogR) [10] and intersection screening (IS) [13, 14] estimators, $M = \mathcal{O}(\log N)$ samples suffice for an Ising model with N spins under certain assumptions, which is consistent with respect to (w.r.t.) previously established information-theoretic lower-bound [11]. The ℓ_1-LogR and IS are both nonlinear M-estimators which minimize the ℓ_1-regularized logistic and IS objective (ISO) loss functions, respectively.

In this paper, we focus on one simpler linear estimator called ℓ_1-regularized linear regression (ℓ_1-LinR) and theoretically investigate its typical learning performance using the replica method [17, 18, 19] from statistical mechanics. The ℓ_1-LinR estimator, also widely known as least absolute shrinkage and selection operator (LASSO) [20] in statistics and machine learning, is considered here mainly for two reasons. On the one hand, it is one representative case of model misspecification since the quadratic loss of ℓ_1-LinR does not match the true log-conditional-likelihood as ℓ_1-LogR, nor does it have the intersection screening property as IS. On the other hand, as one of the most popular linear

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

estimator, ℓ_1-LinR is more computationally efficient than ℓ_1-LogR and IS, and thus it is of practical importance to theoretically investigate its typical performance for Ising model selection. Since it is difficult to obtain the learning performance for general graphs, as a first step we consider the random regular (RR) graphs $\mathcal{G}_{N,d,K_0}$ in the paramagnetic phase [19], where $\mathcal{G}_{N,d,K_0}$ denotes the ensemble of RR graphs with constant node degree d and uniform coupling strength K_0.

1.1 Contributions

The main contributions are as follows. First, we obtain an accurate estimate of the "typical" sample complexity of ℓ_1-LinR for Ising model selection for typical RR graphs in the paramagnetic phase. Specifically, for a typical RR graph $G \in \mathcal{G}_{N,d,K_0}$, using ℓ_1-LinR with a regularization parameter $0 < \lambda < \tanh(K_0)$, one can consistently reconstruct the structure of Ising model as $N \to \infty$ if the number of samples M satisfies $M > \frac{c(\lambda,K_0)\log N}{\lambda^2}$ where $c(\lambda, K_0) = \frac{2\left(1-\tanh^2(K_0)+d\lambda^2\right)}{1+(d-1)\tanh^2(K_0)}$. The accuracy of the obtained "typical" sample complexity is verified by a good agreement between experimental results and theoretical predictions. To the best of our knowledge, this is the first result that provides an accurate "typical" sample complexity even for a limited class of Ising models. Interestingly, as $\lambda \to \tanh(K_0)$, a lower bound $M > \frac{2\log N}{\tanh^2(K_0)}$ of the typical sample complexity is obtained and it has the same scaling as the information-theoretic worst-case lower bound $M > \frac{c' \log N}{K_0^2}$ [11] for some constant c' at high temperatures since $\tanh(K_0) = \mathcal{O}(K_0)$ as $K_0 \to 0$.

Second, we provide a computationally efficient method to precisely predict the typical learning performance of ℓ_1-LinR in the non-asymptotic case with moderate M and N, such as the precision rate, recall rate, and residual sum of square (RSS). Such kind of precise non-asymptotic predictions have not been previously obtained even for ℓ_1-LogR/IS, and are quite different from former asymptotic results of ℓ_1-LinR which assumes fixed ratio $\alpha \equiv M/N$ [21, 22, 23, 24]. Moreover, although our theoretical analysis is based on the tree-like structure assumption, experimental results on two dimensional (2D) grid graphs also show a fairly good agreement, indicating that our theoretical result can be a good approximation even for graphs with many loops.

Third, while this paper focuses on ℓ_1-LinR, our method is readily applicable to a wide class of ℓ_1-regularized estimators, including ℓ_1-LogR [10] and IS [13, 14]. Thus, an additional technical contribution is to provide a generic approach for precisely investigating the typical learning performances of various ℓ_1-regularized estimators for Ising model selection. Although the replica method from statistical mechanics is a non-rigorous method, our result is conjectured to be correct, which is supported by the excellent agreement between experimental results and theoretical predictions.

Finally, our theoretical analysis in this paper provides several useful technical advances which might be of particular interest to the statistical mechanics community. For example, we propose to compute the entropy term by averaging over the Haar measure, as stated in Ansatz (A2), which is otherwise intractable [25]. In addition, in computing the free energy using the replica method, the system dimension N is explicitly kept and thus the finite size effect is taken into account, which is rarely done in previous statistical mechanics analysis [19].

1.2 Related works

There has been some earlier works on the analysis of Ising model selection (also known as the inverse Ising problem) using the replica method [4, 5, 6, 7, 25] from statistical mechanics. For example, in [6], the performance of the pseudo-likelihood (PL) method [26] is studied. However, instead of graph structure learning, [6] focuses on the problem of parameter learning. Then, [7] extends the analysis to the Ising model with sparse couplings using logistic regression without regularization. The recent work [25] analyzes the performance of ℓ_2-regularized linear regression but the techniques invented there are not applicable to ℓ_1-LinR since the ℓ_1-norm breaks the rotational invariance property.

Regarding the study of ℓ_1-LinR (LASSO) under model misspecification, the past few years have seen a line of research in the field of signal processing with a specific focus on the single-index model [27, 28, 23, 29, 30]. These studies are closely related to ours but there are 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 ℓ_1-LinR while most previous studies consider estimation consistency except [29]. However, [29] only considers the classical asymptotic regime while we are interested in the high-dimensional setting where $M \ll N$.

As far as we have searched, there is no earlier study of ℓ_1-LinR estimator for Ising model selection, though some are found for Gaussian graphical models [31, 32]. One closely related work [14] states that at high temperatures when the coupling magnitude is approaching zero, both logistic and IOS losses can be approximated by a square loss. However, their claim is only restricted to the small magnitude near zero while our analysis extends the validity range to the whole paramagnetic phase. Moreover, they evaluate the minimum number of samples necessary for consistently reconstructing "arbitrary" Ising models, which, however, seems much larger than that actually needed. By contrast, we provide the first accurate assessment of sample complexity for consistently reconstructing "typical" samples of Ising models defined over the RR graphs. Furthermore, [14] does not provide precise predictions of the non-asymptotic learning performance as we do.

2 Background and Problem Setup

2.1 Ising Model

Ising model is one special class of MRFs with pairwise potentials and each variable takes binary values [18, 19], which is one classical model from statistical physics. The joint probability distribution of an Ising model with N variables (spins) $\boldsymbol{s} = (s_i)_{i=0}^{N-1} \in \{-1, +1\}^N$ has the form

$$P_{\text{Ising}}(\boldsymbol{s}|\boldsymbol{J}) = \frac{1}{Z(\boldsymbol{J})} \exp\left\{\sum_{i<j} J_{ij} s_i s_j\right\}, \tag{1}$$

where $Z(\boldsymbol{J}) = \sum_{\boldsymbol{s}} \exp\left\{\sum_{i<j} J_{ij} s_i s_j\right\}$ is the partition function and $\boldsymbol{J} = (J_{ij})_{i,j}$ are the couplings, respectively. In general, there are also external fields but here they are assumed to be zero for simplicity. The structure of Ising model can be described by an undirected graph $G = (\mathtt{V}, \mathtt{E})$, where $\mathtt{V} = \{0, 1, ..., N-1\}$ is a collection of vertices at which the spins are assigned, and $\mathtt{E} = \{(i,j) \,|J_{ij} \neq 0\}$ is a collection of undirected edges, i.e., $J_{ij} = 0$ for all pairs of $(i,j) \notin \mathtt{E}$. For each vertex $i \in \mathtt{V}$, its neighborhood is defined as the subset $\mathcal{N}(i) \equiv \{j \in \mathtt{V} | \, (i,j) \in \mathtt{E}\}$.

2.2 Neighborhood-based ℓ_1-regularized linear regression (ℓ_1-LinR)

The problem of Ising model selection refers to recovering the graph G (edge set $\mathtt{E}$), given M i.i.d. samples $\mathcal{D}^M = \left\{\boldsymbol{s}^{(1)}, ..., \boldsymbol{s}^{(M)}\right\}$ from the Ising model. While the maximum likelihood method has nice properties of consistency and asymptotic efficiency, it suffers from high computational complexity. To tackle this difficulty, several local learning algorithms have been proposed, notably the ℓ_1-LogR estimator [10] and IS estimator [13]. Both ℓ_1-LogR and IS optimize a regularized local cost function $\ell(\cdot)$ for each spin i.e., $\forall i \in \mathtt{V}$,

$$\hat{\boldsymbol{J}}_{\setminus i} = \arg\min_{\boldsymbol{J}_{\setminus i}} \left[\frac{1}{M} \sum_{\mu=1}^{M} \ell\left(s_i^{(\mu)} h_{\setminus i}^{(\mu)}\right) + \lambda \left\|\boldsymbol{J}_{\setminus i}\right\|_1\right], \tag{2}$$

where $h_{\setminus i}^{(\mu)} \equiv \sum_{j \neq i} J_{ij} s_j^{(\mu)}$, $\boldsymbol{J}_{\setminus i} \equiv (J_{ij})_{j(\neq i)}$. Specifically, $\ell(x) = \log\left(1 + e^{-2x}\right)$ for ℓ_1-LogR and $\ell(x) = e^{-x}$ for IS, which correspond to the minus log conditional distribution [10] and the ISO [13], respectively. Consequently, the problem of recovering the edge set $\mathtt{E}$ is equivalently reduced to local neighborhood selection, i.e., recovering the neighborhood set $\mathcal{N}(i)$ for each vertex $i \in \mathtt{V}$. In particular, given the estimates $\hat{\boldsymbol{J}}_{\setminus i}$ in (2), the neighborhood set of vertex i can be estimated via the nonzero coefficients, i.e.,

$$\hat{\mathcal{N}}(i) = \left\{j | \hat{J}_{ij} \neq 0, j \in \mathtt{V} \setminus i\right\}, \ \forall i \in \mathtt{V}. \tag{3}$$

In this paper, we focus on one simple linear estimator, termed as the ℓ_1-LinR estimator, i.e., $\forall i \in \mathtt{V}$,

$$\hat{\boldsymbol{J}}_{\setminus i} = \arg\min_{\boldsymbol{J}_{\setminus i}} \left[\frac{1}{2M} \sum_{\mu=1}^{M} \left(s_i^{(\mu)} - h_{\setminus i}^{(\mu)}\right)^2 + \lambda \left\|\boldsymbol{J}_{\setminus i}\right\|_1\right], \tag{4}$$

which, recalling that $s_i^{(\mu)} \in \{-1, +1\}$, corresponds to a square loss $\ell(x) = \frac{1}{2}(x-1)^2$ in (2). The neighborbood set for each vertex $i \in \mathtt{V}$ is estimated in the same way as (3). Interestingly, the square loss used in (4) implies that the postulated conditional distribution is Gaussian and thus inconsistent with the true one, which is one typical case of model misspecification. Furthermore, compared with nonlinear estimators ℓ_1-LogR and IS, the ℓ_1-LinR estimator is more efficient to implement.

3 Statistical Mechanics Analysis

In this section, the statistical mechanics analysis of the ℓ_1-LinR estimator for typical RR graphs in the paramagnetic phase is presented. For simplicity and without loss of generality, we focus on spin s_0 and will drop certain subscript for notational convenience. Following the terminology in [7, 25], we will refer to the Ising model generating the dataset $\mathcal{D}^M$ with couplings $\boldsymbol{J}^*$ as the teacher model. To characterize the structure learning performance of the estimator, the precision and recall are considered:

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}, \tag{5}$$

where TP, FP, FN denote the number of true positive, false positive, and false negative samples in the estimator $\hat{\boldsymbol{J}}$, respectively. Moreover, the residual sum of squares (RSS), defined as $RSS = \left\|\hat{\boldsymbol{J}} - \boldsymbol{J}^*\right\|_2^2$, can also be considered as a metric for parameter learning.

3.1 Problem Formulation

The basic idea of the statistical mechanical approach is to introduce the following Hamiltonian and Boltzmann distribution induced by the loss function $\ell(\cdot)$

$$\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right) = \sum_{\mu=1}^{M} \ell\left(s_0^{(\mu)} h^{(\mu)}\right) + \lambda M \left\|\boldsymbol{J}\right\|_1, \tag{6}$$

$$P\left(\boldsymbol{J}|\mathcal{D}^M\right) = \frac{1}{Z} e^{-\beta \mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)}, \tag{7}$$

where $Z = \int d\boldsymbol{J} e^{-\beta \mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)}$ is the partition function, and $\beta\,(>0)$ is the inverse temperature. For a typical RR graph, in the zero-temperature limit $\beta \to +\infty$, the Boltzmann distribution (7) converges to a point-wise measure on the estimator (2). In statistical mechanics, the macroscopic properties of (7) can be analyzed by assessing the free energy density $f(\mathcal{D}^M) = -\frac{1}{N\beta}\log Z$, which, in the current case, depends on the predetermined randomness of $\mathcal{D}^M$. However, as $N, M \to \infty$, $f(\mathcal{D}^M)$ is expected to show the *self averaging property* [17]: for typical datasets $\mathcal{D}^M$, $f(\mathcal{D}^M)$ converges to its average over the random data $\mathcal{D}^M$:

$$f = -\frac{1}{N\beta}\left[\log Z\right]_{\mathcal{D}^M}, \tag{8}$$

where $[\cdot]_{\mathcal{D}^M}$ denotes expectation over $\mathcal{D}^M$, i.e. $[\cdot]_{\mathcal{D}^M} = \sum_{\boldsymbol{s}^{(1)},\ldots,\boldsymbol{s}^{(M)}}(\cdot)\prod_{\mu=1}^{M} P_{\text{Ising}}\left(\boldsymbol{s}^{(\mu)}|\boldsymbol{J}^*\right)$. Consequently, one can analyze the typical performance of (7) and hence the estimator (4) as a special case of $\ell(x) = \frac{1}{2}(x-1)^2$, via the assessment of (8).

3.2 Replica computation of the free energy density

Unfortunately, computing (8) rigorously is difficult. For practically overcoming this difficulty, we resort to the replica method [17, 18, 19] from statistical mechanics, which is symbolized using the following identity

$$f = -\frac{1}{N\beta}\left[\log Z\right]_{\mathcal{D}^M} = -\lim_{n\to 0}\frac{1}{N\beta}\frac{\partial \log\left[Z^n\right]_{\mathcal{D}^M}}{\partial n}. \tag{9}$$

The basic idea is as follows. One replaces the average of $\log Z$ by that of Z^n which is analytically tractable for $n \in \mathbb{N}$ in the large N limit, and constructs an analytically continuable expression from $\mathbb{N}$ to $\mathbb{R}$, then takes the limit $n \to 0$ by using the expression. Although the replica method is not rigorous, it has been empirically verified from extensive studies in disorder systems [18, 19] and also found useful in the study of high-dimensional models in machine learning [33, 34].

Specifically, with the Hamiltonian $\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)$, assuming $n \in \mathbb{N}$ is a positive integer, the replicated partition function $\left[Z^n\right]_{\mathcal{D}^M}$ in (9) can be written as

$$\left[Z^n\right]_{\mathcal{D}^M} = \int \prod_{a=1}^{n} d\boldsymbol{J}^a e^{-\beta\lambda M \sum_{a=1}^{n}\left\|\boldsymbol{J}^a\right\|_1} \left\{\sum_{\boldsymbol{s}} P_{\text{Ising}}\left(\boldsymbol{s}|\boldsymbol{J}^*\right)\exp\left[-\beta\sum_{a=1}^{n}\ell\left(s_0 h^a\right)\right]\right\}^M, \tag{10}$$

where $h^a = \sum_j J_j^a s_j$ will be termed as *local field* hereafter. The analysis below essentially depends on the distribution of the local field but it is nontrivial. To resolve this problem, we here take the similar approach in [7, 25] and introduce the following ansatz.

Ansatz 1 (A1): *Denote $\Psi = \{j | j \in \mathcal{N}(0)\}$ and $\bar{\Psi} = \{j | j = 1, ..., N-1, j \notin \mathcal{N}(0)\}$ as the active and inactive sets of spin s_0, respectively, then for a typical RR graph $G \in \mathcal{G}_{N,d,K_0}$ in the paramagnetic phase, i.e., $(d-1)\tanh^2(K_0) < 1$, the ℓ_1-LinR estimator in (4) obeys the following form*

$$\hat{J}_j = \begin{cases} \bar{J}_j + \frac{1}{\sqrt{N}} w_j, & j \in \Psi \\ \frac{1}{\sqrt{N}} w_j, & j \in \bar{\Psi} \end{cases} \tag{11}$$

where $\bar{J}_j$ is the mean value of the estimator and w_j is a random variable which is asymptotically zero mean with variance scaled as $\mathcal{O}(1)$.

The consistency of Ansatz 1 is checked in Appendix B. Under Ansatz 1, the local fields h^a can be decomposed as $h^a = \sum_{j \in \Psi} \bar{J}_j s_j + h_w^a$ where $h_w^a \equiv \sum_j \frac{1}{\sqrt{N}} w_i^a s_j$ is the "noise" part. According to the central limit theorem (CLT), h_w^a can be approximated as multivariate Gaussian variables, which, under the replica symmetric (RS) ansatz [17], can be fully described by two order parameters

$$Q \equiv \frac{1}{N} \sum_{i,j} w_i^a C_{ij}^{\backslash 0} w_j^a, \quad q \equiv \frac{1}{N} \sum_{i,j} w_i^a C_{ij}^{\backslash 0} w_j^b, (a \neq b), \tag{12}$$

where $C^{\backslash 0} \equiv \{C_{ij}^{\backslash 0}\}$ is the covariance matrix of the teacher Ising model without the spin s_0. Since the difference between $C^{\backslash 0}$ and that with s_0 is not essential in the limit $N \to \infty$, hereafter the superscript $^{\backslash 0}$ will be discarded. As shown in Appendix A, the average free energy density (9) in the limit $\beta \to \infty$ can be computed as

$$f(\beta \to \infty) = -\texttt{Extr}\{-\xi + S\}, \tag{13}$$

where ξ, S are the corresponding energy and entropy terms:

$$S = \lim_{n \to 0} \frac{1}{N\beta} \frac{\partial}{\partial n} \log I, \tag{14}$$

$$I = \int \prod_{a=1}^{n} dw^a \prod_{a=1}^{n} e^{-\lambda\beta \|w^a\|_1} \delta\left(\sum_{i,j} w_i^a C_{ij} w_j^a - NQ\right) \times \prod_{a<b} \delta\left(\sum_{i,j} w_i^a C_{ij} w_j^b - Nq\right), \tag{15}$$

$$\xi = \frac{\alpha \mathbb{E}_{s,z}\left(s_0 - \sum_{j \in \Psi} \bar{J}_j s_j - \sqrt{Q} z\right)^2}{2(1+\chi)} + \alpha\lambda \sum_{j \in \Psi} \left|\bar{J}_j\right|, \tag{16}$$

where $\alpha \equiv M/N, \chi \equiv \lim_{\beta \to \infty} \beta(Q - q)$, $\mathbb{E}_{s,z}(\cdot)$ denotes the expectation operation w.r.t. $z \sim \mathcal{N}(0,1)$ on top of $(s_0, \boldsymbol{s}_\Psi) \sim P_{\text{Ising}}(s_0, \boldsymbol{s}_\Psi | \boldsymbol{J}^*) \propto e^{s_0 \sum_{j \in \Psi} J_j^* s_j}$ [7], and $\texttt{Extr}\{\cdot\}$ denotes the extremum operation w.r.t. relevant variables.

In contrast to the case of ℓ_2-norm in [25], the ℓ_1-norm in (15) breaks the rotational invariance property, i.e., $\|w^a\|_1 \neq \|Ow^a\|_1$ for general orthogonal matrix O, making it difficult to compute the entropy term S. To circumvent this difficulty, we employ an observation that, when considering the RR graph ensemble $\mathcal{G}_{N,d,K_0}$ as the coupling network of the Ising model, the orthogonal matrix O diagonalizing the covariance matrix C appears to be distributed from the Haar orthogonal measure [35, 36]. Thus, it is assumed that I in (15) can be replaced by its average $[I]_O$ over the Haar-distributed O:

Ansatz 2 (A2): *Denote $C \equiv \mathbb{E}_{\boldsymbol{s}}[\boldsymbol{s}\boldsymbol{s}^T]$, where $\mathbb{E}_{\boldsymbol{s}}[\cdot] = \sum_{\boldsymbol{s}} P_{\text{Ising}}(\boldsymbol{s}|\boldsymbol{J}^*)(\cdot)$, as the covariance matrix of spin configurations $\boldsymbol{s}$. Suppose that the eigendecomposition of C is $C = O\Lambda O^T$, where O is the orthogonal matrix, then O can be seen as a random sample generated from the Haar orthogonal measure and thus for typical graph realizations from $\mathcal{G}_{N,d,K_0}$, I in (15) is equal to the average $[I]_O$.*

The consistency of Ansatz (A2) is partly checked in Appendix C. Under Ansatz (A2), the entropy term S in (14) can be alternatively computed as $\lim_{n \to 0} \frac{1}{N\beta} \frac{\partial}{\partial n} \log [I]_O$, as shown in Appendix A. Finally,

under the RS ansatz, the average free energy density (9) in the limit $\beta \to \infty$ is calculated to be

$$f(\beta \to \infty) = -\underset{\Theta}{\text{Extr}} \left\{ \begin{array}{c} -\frac{\alpha}{2(1+\chi)} \mathbb{E}_{s,z} \left(\left(s_0 - \sum_{j \in \Psi} \bar{J}_j s_j - \sqrt{Q} z \right)^2 \right) - \lambda \alpha \sum_{j \in \Psi} \left| \bar{J}_j \right| \\ + (-ER + F\eta) G'(-E\eta) + \frac{1}{2} EQ - \frac{1}{2} F\chi + \frac{1}{2} KR - \frac{1}{2} H\eta \\ -\mathbb{E}_z \min_w \left\{ \frac{K}{2} w^2 - \sqrt{H} z w + \frac{\lambda M}{\sqrt{N}} |w| \right\} \end{array} \right\}, \tag{17}$$

where $z \sim \mathcal{N}(0,1)$, and $G(x)$ is a function defined as

$$G(x) = -\frac{1}{2} \log x - \frac{1}{2} + \underset{\Lambda}{\texttt{Extr}} \left\{ -\frac{1}{2} \int \log(\Lambda - \gamma) \rho(\gamma) d\gamma + \frac{\Lambda}{2} x \right\}, \tag{18}$$

and $\rho(\gamma)$ is the eigenvalue distribution (EVD) of the covariance matrix C, and Θ is a collection of macroscopic parameters $\Theta = \left\{ \chi, Q, E, R, F, \eta, K, H, \left\{ \bar{J}_j \right\}_{j \in \Psi} \right\}$. For details of these macroscopic parameters and $\rho(\gamma)$, please refer to Appendix A and F, respectively. Although there are no analytic solutions, these macroscopic parameters can be obtained by numerically solving the following equations which are termed hereafter as equations of state (EOS) employing the physics terminology:

$$\begin{cases} E = \frac{\alpha}{(1+\chi)}, \\ F = \frac{\alpha}{(1+\chi)^2} \left[\mathbb{E}_s \left(s_0 - \sum_{j \in \Psi} s_j \bar{J}_j \right)^2 + Q \right], \\ R = \frac{1}{K^2} [\left(H + \frac{\lambda^2 M^2}{N} \right) \text{erfc} \left(\frac{\lambda M}{\sqrt{2HN}} \right) - 2\lambda M \sqrt{\frac{H}{N}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\lambda^2 M^2}{2HN}}], \\ E\eta = -\int \frac{\rho(\gamma)}{\tilde{\Lambda} - \gamma} d\gamma, \\ Q = \frac{F}{E^2} + R\tilde{\Lambda} - \frac{(-ER + F\eta)\eta}{\int \frac{\rho(\gamma)}{(\tilde{\Lambda} - \gamma)^2} d\gamma}, \\ K = E\tilde{\Lambda} + \frac{1}{\eta}, \\ \chi = \frac{1}{E} + \eta \tilde{\Lambda}, \\ H = \frac{R}{\eta^2} + F\tilde{\Lambda} + \frac{(-ER + F\eta)E}{\int \frac{\rho(\gamma)}{(\tilde{\Lambda} - \gamma)^2} d\gamma}, \\ \eta = \frac{1}{K} \text{erfc} \left(\frac{\lambda M}{\sqrt{2HN}} \right), \\ \bar{J}_j = \frac{\texttt{soft}(\tanh(K_0), \lambda(1+\chi))}{1 + (d-1)\tanh^2(K_0)}, j \in \Psi, \end{cases} \tag{19}$$

where $\tilde{\Lambda}$ satisfying $E\eta = -\int \frac{\rho(\gamma)}{\tilde{\Lambda} - \gamma} d\gamma$ is determined by the extremization condition in (18) and $\texttt{soft}(z, \tau) = \texttt{sign}(z)(|z| - \tau)_+$ is the soft-thresholding function. Note that in (17), apart from the ratio $\alpha \equiv M/N$, N and M also appear as $\lambda M/\sqrt{N}$ in the free energy result, which is different from previous results [7, 25, 33]. The reason is that, thanks to the ℓ_1-regularization term $\lambda M \|\boldsymbol{J}\|_1$ in the Hamiltonian $\mathcal{H}\left(\boldsymbol{J} | \mathcal{D}^M\right)$ (6), the mean estimates $\left\{ \bar{J}_j \right\}_{j \in \Psi}$ in the active set Ψ and the noise w in the inactive set $\bar{\Psi}$ essentially give different scaling contributions to the free energy density, i.e., $\lambda\alpha$ before $\sum_{j \in \Psi} \left| \bar{J}_j \right|$ and $\lambda M/\sqrt{N}$ before $|w|$ in (17). Consequently, the two different scaling factors cannot be simultaneously absorbed by any scaling change of λ.

3.3 High-dimensional asymptotic result

From the free energy result (17) and Ansatz (A1) (11), we obtain explicit expressions of the ℓ_1-LinR estimator. Specifically, the mean estimates $\{\bar{J}_j\}_{j \in \Psi}$ in the active set Ψ are

$$\bar{J}_j = \frac{\texttt{soft}(\tanh(K_0), \lambda(1+\chi))}{1 + (d-1)\tanh^2(K_0)}, j \in \Psi, \tag{20}$$

while the noise estimates $\{\hat{J}_j\}_{j \in \bar{\Psi}}$ in the inactive set $\bar{\Psi}$ are

$$\hat{J}_j = \frac{\sqrt{H}}{K\sqrt{N}} \texttt{soft}\left(z_j, \frac{\lambda M}{\sqrt{HN}} \right), j \in \bar{\Psi}, \tag{21}$$

where $z_j \sim \mathcal{N}(0,1), j \in \bar{\Psi}$ are i.i.d. standard Gaussian random variables. The results (20) and (21) assert that the ℓ_1-LinR estimator is decoupled and its asymptotic behavior can be described by two

scalar soft-thresholding estimators for the active set and inactive set, respectively. Consequently, the statistical properties of ℓ_1-LinR can be readily obtained once the EOS (19) is solved. The derivation and interpretation of (20) and (21) are in Appendix A.3.

In the high-dimensional setting where the number of vertices N is allowed to grow as a function of the number of samples M, one important question for Ising model selection is that what is the minimum number of samples M required to consistently recover the structure as $N \to \infty$. In particular, the ℓ_1-LinR estimator (4) is called *model selection consistent* if both the associated precision and recall satisfy $Precision \to 1$ and $Recall \to 1$ as $N \to \infty$. Though it can be evaluated with the two scalar estimators (20) and (21), there are no analytical solutions to EOS (19), which makes it difficult to derive an explicit condition. To overcome this difficulty, as shown in Appendix D, we perform perturbation analysis of EOS (17) and obtain an asymptotic relation $H \simeq F$, Then, we obtain that for a RR graph $G \in \mathcal{G}_{N,d,K_0}$, given M i.i.d. samples $\mathcal{D}^M$, the ℓ_1-LinR estimator (4) can successfully recover the graph structure G as $N \to \infty$ if

$$M > \frac{c\left(\lambda, K_0\right) \log N}{\lambda^2}, 0 < \lambda < \tanh\left(K_0\right), \tag{22}$$

where $c(\lambda, K_0)$ is a constant value dependent on the regularization parameter λ and coupling strength K_0 and a sharp prediction (as verified in Sec. 4) is obtained as

$$c\left(\lambda, K_0\right) = \frac{2\left(1 - \tanh^2\left(K_0\right) + d\lambda^2\right)}{1 + (d-1)\tanh^2\left(K_0\right)}. \tag{23}$$

For details of the analysis, including that of ℓ_1-LogR, see Appendix D. Consequently, we obtain the "typical" sample complexity of ℓ_1-LinR for Ising model selection for RR graphs in the paramagnetic phase. The result in (22) is derived for ℓ_1-LinR with a fixed regularization parameter λ. Since the value of λ is upper bounded by $\tanh\left(K_0\right)$ (otherwise false negatives occur as discussed in Appendix D), a lower bound of the typical sample complexity for ℓ_1-LinR is obtained as

$$M > \frac{2 \log N}{\tanh^2\left(K_0\right)}. \tag{24}$$

Interestingly, the scaling of the typical result in (24) is the same as the information-theoretic worst-case result $M > \frac{c \log N}{K_0^2}$ obtained in [11] at high temperatures since $\tanh\left(K_0\right) = \mathcal{O}\left(K_0\right)$ as $K_0 \to 0$.

3.4 Non-asymptotic result for moderate M, N

It is desirable in practice to predict the non-asymptotic performance of the ℓ_1-LinR estimator for finite M, N. However, it is found that solving (20) and (21) simply by inserting finite values of M, N does not provide good consistency with some of the experimental results, in particular the recall rate. This is reasonable since in obtaining the energy term ξ in (16), the fluctuations around the mean estimates $\left\{\bar{J}_j\right\}_{j \in \Psi}$ due to finite size effect of M is not taken into account properly by the expectation $\mathbb{E}_{s,z}\left(\cdot\right)$. To address this problem, we replace $\mathbb{E}_{s,z}\left(\cdot\right)$ in (16) with finite sample average, whereby the estimates $\{\hat{J}_j\}_{j \in \Psi}$ can be obtained by solving the following d-dimensional optimization problem

$$\min_{J_{j, j \in \Psi}} \left[\frac{\sum_{\mu=1}^{M} \left(s_0^\mu - \sum_{j \in \Psi} s_j^\mu J_j - \sqrt{Q} z^\mu\right)^2}{2\left(1+\chi\right) M} + \lambda \sum_{j \in \Psi} \left|J_j\right| \right], \tag{25}$$

where $s_0^\mu, s_{j, j \in \Psi}^\mu \sim P\left(s_0, \boldsymbol{s}_\Psi | \boldsymbol{J}^*\right),\ z^\mu \sim \mathcal{N}\left(0,1\right), \mu = 1 ... M$. The solution to (25) is equivalent to (20) as $M \to \infty$ but (25) enables us to capture the fluctuations of $\{\hat{J}_j\}_{j \in \Psi}$ when M is finite. Note that due to the modification in (25), the solutions to the EOS (19) also need to be modified accordingly to take into account the finite size effect. One can solve them iteratively, as sketched in Algorithm 1 and the implementation details are shown in Appendix E.1.

Consequently, given the solutions of R, H, Q, χ in Algorithm 1 for moderate M, N, the non-asymptotic statistical properties of the ℓ_1-LinR estimator can be evaluated using computationally tractable MC simulations to the reduced d-dimensional ℓ_1-LinR estimator (25) and scalar estimator (21). Denote $\{\hat{J}_j^t\}, t = 1, ..., T_{\mathrm{MC}}$ as the estimates in t-th MC simulation, where $\hat{J}_{j, j \in \Psi}^t$ and $\hat{J}_{j, j \in \bar{\Psi}}^t$

Algorithm 1 Method to solve EOS (19) together with (25) for moderate M, N.

Initialization: $\chi, Q, E, R, F, \eta, K, H$.
repeat
  for $t = 1$ **to** $T_{\rm MC}$ **do**
    Draw random samples $s_0^\mu, s_{j,j\in\Psi}^\mu \sim P\left(s_0, \boldsymbol{s}_\Psi | \boldsymbol{J}^*\right)$ and $z^\mu \sim \mathcal{N}(0,1)$, $\mu = 1...M$.
    Obtain solutions $\left\{\hat{J}_j\right\}_{j\in\Psi}$ to (25).
    Compute $\triangle(t) = \frac{1}{M}\sum_{\mu=1}^{M}\left(s_0^\mu - \sum_{j\in\Psi} s_j^\mu \hat{J}_j\right)^2$.
  end for
  Update the values of $\chi, Q, E, R, F, \eta, K, H$ by solving the EOS (19) with the substitution of $\mathbb{E}_s\left(s_i - \sum_{j\in\Psi} s_j \bar{J}_j\right)^2 = \frac{1}{T_{\rm MC}}\sum_{t=1}^{T_{\rm MC}} \triangle(t)$.
until convergence.

are solutions of (25) and (21), and $T_{\rm MC}$ is the total number of MC simulations. Then, the $Precision$ and $Recall$ are computed as

$$Precision = \frac{1}{T_{\rm MC}}\sum_{t=1}^{T_{\rm MC}} \frac{\left\|\hat{J}^t_{j,j\in\Psi}\right\|_0}{\left\|\hat{J}^t_{j,j\in\Psi}\right\|_0 + \left\|\hat{J}^t_{j,j\in\bar{\Psi}}\right\|_0}, \quad Recall = \frac{1}{T_{\rm MC}}\sum_{t=1}^{T_{\rm MC}} \frac{\left\|\hat{J}^t_{j,j\in\Psi}\right\|_0}{d},, \tag{26}$$

where $\|\cdot\|_0$ is the ℓ_0-norm indicating the number of nonzero elements. In addition, the RSS can be computed as $RSS = \frac{1}{T_{\rm MC}}\sum_{t=1}^{T_{\rm MC}}\sum_{j\in\Psi}\left|\hat{J}^t_j - K_0\right|^2 + R$.

4 Experimental results

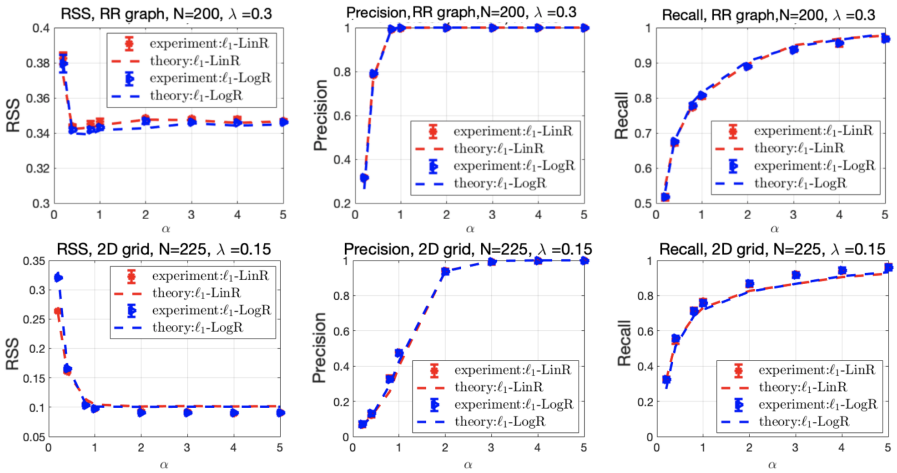

Figure 1: Theoretical and experimental results of RSS, $Precision$ and $Recall$ for RR graph and 2D grid using ℓ_1-LinR and ℓ_1-LogR with different values of $\alpha \equiv M/N$. The standard error bars are obtained from 5 random runs, each with 10^3 MC simulations. A good agreement between theory and experiment is achieved, even for small-size 2D grid graph with many loops.

In this section, we conduct numerical experiments to verify the accuracy of the theoretical analysis. The experimental procedures are as follows. First, a random graph $G \in \mathcal{G}_{N,d,K_0}$ is generated and the Ising model is defined on it. Then, the spin snapshots are obtained using MC sampling, yielding the dataset $\mathcal{D}^M$. We randomly choose a center spin s_0 and infer its neighborhood using the ℓ_1-LinR (4) and ℓ_1-LogR [10] estimators. To obtain standard error bars, we repeat the sequence of operations many times. The RR graph $G \in \mathcal{G}_{N,d,K_0}$ with node degree $d = 3$ and coupling strength $K_0 = 0.4$ is

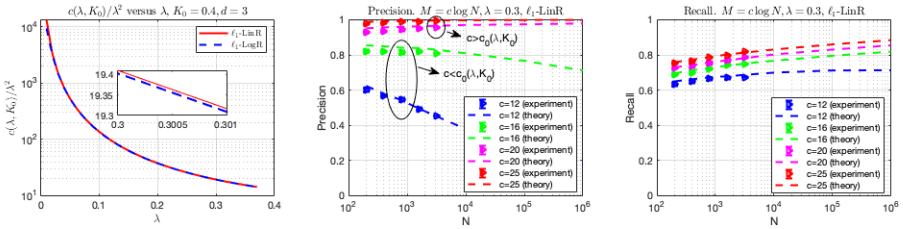

Figure 2: Left: Critical scaling value $c_0(\lambda, K_0) \equiv \frac{c(\lambda,K_0)}{\lambda^2}$ of ℓ_1-LinR and ℓ_1-LogR for the RR graph $G \in \mathcal{G}_{N,d,K_0}$ with $d = 3, K_0 = 0.4$. Middle and Right: Precision and Recall for RR graph using ℓ_1-LinR with $\lambda = 0.3$. Experimental results are shown for $N = 200, 400, 800, 1600, 3200$. When $c > c_0(\lambda, K_0)$ ($c_0(\lambda = 0.3, K_0) \approx 19.41$ in this case), the *Precision* increases consistently with N and approaches 1 as $N \to \infty$ while it decreases consistently with N when $c < c_0(\lambda, K_0)$. The *Recall* increases consistently and approach to 1 as $N \to \infty$.

considered, which satisfies the paramagnetic condition $(d-1)\tanh^2(K_0) < 1$. The active couplings $\{J_{ij}\}_{(i,j)\in \mathrm{E}}$ have the same probability of taking both signs of $+1$ or -1 [1].

We first verify the precise non-asymptotic predictions of our method described in Sec.3.4. Fig. 1 (upper figure) shows the replica and experimental results of RSS, $Precision$, $Recall$ for $N = 200$ with different values of $\alpha \equiv M/N$. It can be seen that for both ℓ_1-LinR and ℓ_1-LogR, there is a fairly good agreement between the theoretical predictions and experimental results, even for small $N = 200$ and small α (equivalently small M), verifying the correctness of the replica analysis. Interestingly, the quantitatively similar behavior between ℓ_1-LinR and ℓ_1-LogR is observed. Note that our theoretical analysis is based on the tree-like structure assumption and hence it is interesting to see if it is applicable to graphs with loops. To this end, we consider the 2D 4-nearest neighbor grid with periodic boundary condition, which is one common loopy graph. Fig. 1 (lower figure) shows the results for a 15×15 2D grid with uniform constant coupling $K_0 = 0.2$. The agreement between the theoretical and numerical results is fairly good, indicating that our theoretical result can be a good approximation even for loopy graphs. More results for different values of N and λ are shown in Fig. 6 and Fig. 7 in Appendix G.

Subsequently, the asymptotic result and sharpness of the critical scaling value $c_0(\lambda, K_0) \equiv \frac{c(\lambda,K_0)}{\lambda^2}$ in (22) are evaluated. First, Fig. 2 (left) shows comparison of $c_0(\lambda, K_0)$ between ℓ_1-LinR and ℓ_1-LogR for the RR graph $G \in \mathcal{G}_{N,d,K_0}$ when $d = 3, K_0 = 0.4$, indicating similar behavior of ℓ_1-LogR and ℓ_1-LinR. Then, we conduct experiments for $M = c \log N$ with different values of c around $c_0(\lambda, K_0)$, and investigate the trend of $Precision$ and $Recall$ as N increases. When $\lambda = 0.3$, Fig. 2 (middle and right) show the results of *Precision* and *Recall*, respectively. As expected, the *Precision* increases consistently with N when $c > c_0(\lambda, K_0)$ and decreases consistently with N when $c < c_0(\lambda, K_0)$ while the *Recall* increases consistently and approaches to 1 as $N \to \infty$, which verifies the sharpness of the critical scaling value prediction. The results for ℓ_1-LogR, including the case of $\lambda = 0.1$ for both ℓ_1-LinR and ℓ_1-LogR, are shown in Fig. 8 and Fig. 9 in Appendix G.

5 Conclusion

In this paper, we provide a statistical mechanics analysis of the typical learning performance of ℓ_1-LinR for Ising model selection. An accurate estimate of the typical sample complexity of ℓ_1-LinR is derived for typical RR graphs in the paramagnetic phase. It is demonstrated that ℓ_1-LinR, albeit with model misspecification, is model selection consistent with $M = \mathcal{O}(\log N)$ samples, which is of the same order as ℓ_1-LogR. Moreover, application of the replica method from statistical mechanics allows us to obtain sharp predictions of the non-asymptotic behaviour of ℓ_1-LinR for moderate size of M, N. There is an excellent agreement between the theoretical predictions and experimental results, even for graphs with many loops, which supports our findings. Several key assumptions are made in our theoretical analysis for the Ising model, such as the paramagnetic assumption which implies that the coupling strength should be not too large. These assumptions restrict the applicability of the presented result, and thus overcoming such limitations will be an important direction for future work.

[1] Though this setting is different from the analysis where the nonzero teacher couplings take a uniform sign, the result can be directly compared thanks to gauge symmetry [17].

References

[1] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[2] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017.

[3] Erik Aurell and Magnus Ekeberg. Inverse ising inference using all the data. *Physical review letters*, 108(9):090201, 2012.

[4] Ludovica Bachschmid-Romano and Manfred Opper. Learning of couplings for random asymmetric kinetic ising models revisited: random correlation matrices and learning curves. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(9):P09016, 2015.

[5] Johannes Berg. Statistical mechanics of the inverse ising problem and the optimal objective function. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(8):083402, 2017.

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

[7] Alia Abbara, Yoshiyuki Kabashima, Tomoyuki Obuchi, and Yingying Xu. Learning performance in inverse ising problems with sparse teacher couplings. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(7):073402, 2020.

[8] Martin J Wainwright, John D Lafferty, and Pradeep K Ravikumar. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in neural information processing systems*, pages 1465–1472, 2007.

[9] Holger Höfling and Robert Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10(4):883–906, 2009.

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

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

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

[13] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[14] Andrey Y Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. Optimal structure and parameter learning of ising models. *Science advances*, 4(3):e1700791, 2018.

[15] Guy Bresler, Elchanan Mossel, and Allan Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 343–356. Springer, 2008.

[16] Adarsh Prasad, Vishwak Srinivasan, Sivaraman Balakrishnan, and Pradeep Ravikumar. On learning ising models under huber's contamination model. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[18] Manfred Opper and David Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[19] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

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

[21] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[22] Sundeep Rangan, Alyson K Fletcher, and Vivek K Goyal. Asymptotic analysis of map estimation via the replica method and applications to compressed sensing. *IEEE Transactions on Information Theory*, 58(3):1902–1923, 2012.

[23] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, 28:3420–3428, 2015.

[24] Cédric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for convex penalized linear regression beyond gaussian matrices. *arXiv preprint arXiv:2002.04372*, 2020.

[25] Xiangming Meng, Tomoyuki Obuchi, and Yoshiyuki Kabashima. Structure learning in inverse ising problems using ℓ_2-regularized linear estimator. *arXiv preprint arXiv:2008.08342*, 2020.

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

[27] David R Brillinger. A generalized linear model with gaussian regressor variables. In *A Festschrift for Erich L. Lehmann*, page 97–114. 1982.

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

[29] Yue Zhang, Weihong Guo, and Soumya Ray. On the consistency of feature selection with lasso for non-linear targets. In *International Conference on Machine Learning*, pages 183–191. PMLR, 2016.

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

[31] Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

[32] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[33] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

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

[35] Persi Diaconis and Mehrdad Shahshahani. On the eigenvalues of random matrices. *Journal of Applied Probability*, 31(A):49–62, 1994.

[36] Kurt Johansson. On random matrices from the compact classical groups. *Annals of mathematics*, pages 519–545, 1997.

[37] Federico Ricci-Tersenghi. 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, 2012.

[38] H Chau Nguyen and Johannes Berg. Bethe–peierls approximation and the inverse ising problem. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(03):P03004, 2012.

[39] Brendan D McKay. The expected eigenvalue distribution of a large regular graph. *Linear Algebra and its Applications*, 40:203–216, 1981.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.2
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 3 and Appendices A - F.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 4
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]