Sample-Efficient L0-L2 Constrained Structure Learning of Sparse Ising Models

Antoine Dedieu, Miguel Lázaro-Gredilla, Dileep George

Vicarious AI
{antoine, miguel, dileep}@vicarious.com

Abstract

We consider the problem of learning the underlying graph of a sparse Ising model with p nodes from n i.i.d. samples. The most recent and best performing approaches combine an empirical loss (the logistic regression loss or the interaction screening loss) with a regularizer (an L1 penalty or an L1 constraint). This results in a convex problem that can be solved separately for each node of the graph. In this work, we leverage the cardinality constraint L0 norm, which is known to properly induce sparsity, and further combine it with an L2 norm to better model the non-zero coefficients. We show that our proposed estimators achieve an improved sample complexity, both (a) theoretically, by reaching new state-of-the-art upper bounds for recovery guarantees, and (b) empirically, by showing sharper phase transitions between poor and full recovery for graph topologies studied in the literature, when compared to their L1-based state-of-the-art methods.

1 Introduction

Ising models are extremely popular and useful tools (Onsager 1944; McCoy and Maillard 2012) with origins dating back to the $1920s$ (Ising 1925) which have been widely studied in the fields of statistical physics (Thompson 2015) and Bayesian modeling (Bishop 2006). Given an integer p and a symmetric matrix $\boldsymbol{W}^* \in \mathbb{R}^{p \times p}$ with zero diagonal, a binary Ising graph (also known as fully visible Boltzman machine) with p nodes and without external fields specifies the probability of a binary vector $\boldsymbol{z} \in \{-1, 1\}^p$ as

$$p(\boldsymbol{z}|\boldsymbol{W}^*) = \frac{1}{Z(\boldsymbol{W}^*)} \exp\left(\frac{1}{2}\boldsymbol{z}^T\boldsymbol{W}^*\boldsymbol{z}\right), \quad (1)$$

where $Z(\boldsymbol{W}^*)$ is a normalization term called the partition function, which is defined as

$$Z(\boldsymbol{W}^*) = \sum_{\boldsymbol{z} \in \{-1,1\}^p} \exp\left(\frac{1}{2}\boldsymbol{z}^T\boldsymbol{W}^*\boldsymbol{z}\right). \quad (2)$$

We define the connectivity graph associated with $\boldsymbol{W}^*$ as the graph $G = (V, E)$ with vertices $V = \{1, \ldots, p\}$ and edges $E = \{(i, j) : W_{ij}^* \neq 0\}$. We additionally assume that each vertex of G has a degree at most k^* — which is equivalent from saying that each row of $\boldsymbol{W}^*$ has at most k^* nonzero entries — as it is traditionally the case in the literature (Bresler 2015; Vuffray et al. 2016; Lokhov et al. 2018).

In this paper[1], we consider the problem of learning the Ising model in Equation (1) from n samples independently drawn from the model. In particular, we are interested in analyzing the number of samples required for theoretical and empirical recovery of the connectivity graph. Maximum likelihood estimators are intractable for this task (Bishop 2006; Welling and Sutton 2005) as the computational cost of estimating the partition function is exponential with the number of nodes in the graph. While early attempts to learn Ising models were based on mean-field approximation (Tanaka 1998), recent work has shown that the connectivity graph can be efficiently recovered by solving a convex problem for each node, without estimating the partition function. We review the two most recent and influential methods.

Notation: For a vector $\boldsymbol{u} \in \mathbb{R}^p$ and an index $j \in \{1, \ldots, p\}$, we denote $\boldsymbol{u}_{-j} = (u_1, \ldots, u_{j-1}, u_{j+1}, u_p) \in \mathbb{R}^{p-1}$ the vector with all entries except the jth one.

Method 1: Logistic regression estimators For a sample $\boldsymbol{z}$ following Equation (1) and a node $j \in \{1, \ldots, p\}$ of the Ising graph, the conditional probability of the jth observation z_j given all the other observations $\boldsymbol{z}_{-j}$ can be expressed

$$p(z_j|\boldsymbol{z}_{-j}, \boldsymbol{W}^*) = \frac{1}{1 + \exp(-2z_j(\boldsymbol{w}_{-j}^*)^T\boldsymbol{z}_{-j})}, \quad (3)$$

where we have noted $\boldsymbol{w}_{-j}^* = (\boldsymbol{e}_j^T\boldsymbol{W}^*)_{-j} \in \mathbb{R}^{p-1}$ the jth row of the connectivity matrix, without the (null) diagonal term. Consequently, given n independent observations $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}$, the (normalized) pseudo-likelihood (PL) estimator (Besag 1975) of $\boldsymbol{W}^*$ is computed by solving the convex logistic regression (LR) problem for each node:

$$\tilde{\boldsymbol{w}}_{-j} \in \underset{\boldsymbol{w} \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-2y_i\boldsymbol{x}_i^T\boldsymbol{w})), \quad (4)$$

where, for the jth node, we have we denoted $y_i = z_j^{(i)}$ and $\boldsymbol{x}_i = \boldsymbol{z}_{-j}^{(i)}$ for the sake of simplicity. We can then derive an estimator $\tilde{\boldsymbol{W}}$ of $\boldsymbol{W}^*$ by solving Problem (4) for each node of the graph. The symmetrized estimator $\hat{\boldsymbol{W}}$ is then derived by sharing the parameters for both halves (Ravikumar et al.

---

[1]Our code is available at https://github.com/antoine-dedieu/structure_learning_sparse_ising_models

2010; Vuffray et al. 2016; Lokhov et al. 2018), that is by defining $\hat{W}_{ij} = \frac{1}{2}(\tilde{W}_{ij} + \tilde{W}_{ji}), \forall i, j$. When the graph structure is known to be sparse, an influential work (Ravikumar et al. 2010) proposes to add an L1 regularization to encourage sparsity in the coefficients. That is, the authors solve for each node the convex problem

$$\min_{\boldsymbol{w} \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-2y_i \boldsymbol{x}_i^T \boldsymbol{w})) + \lambda \|\boldsymbol{w}\|_1, \quad (5)$$

where the regularization parameter $\lambda \geq 0$ controls the degree of shrinkage on $\boldsymbol{w}$. A well-known limitation of their approach (Montanari and Pereira 2009) is that their theoretical guarantees hold for a small class of models which satisfies restricted isometry properties (RIP) that are NP-hard to validate (Bandeira et al. 2013). To overcome this, a recent work (Wu, Sanghavi, and Dimakis 2019) considers the constrained version of Problem (5) and estimates the sparse connectivity graph by solving:

$$\min_{\boldsymbol{w} \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-2y_i \boldsymbol{x}_i^T \boldsymbol{w})) \quad \text{s.t.} \quad \|\boldsymbol{w}\|_1 \leq \lambda, \tag{6}$$

where λ is selected so that $\lambda \geq \max_{i=1,\dots,n} \|\boldsymbol{w}_{-j}^*\|_1$. The authors additionally derive the best upper bound known on the number of samples required for theoretical graph recovery guarantees without any RIP (cf. Section 3).

Method 2: Interaction screening estimator A recent work (Vuffray et al. 2016) introduces as an alternative the L1-regularized interaction screening estimator (ISE), which is defined at node j as a solution of the convex problem:

$$\min_{\boldsymbol{w} \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \exp(-y_i \mathbf{x}_i^T \boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1. \tag{7}$$

The authors study the sample complexity of ISE for recovery guarantees (cf. Section 3) and additionally show that L1-regularized ISE outperforms L1-regularized LR through a rich variety of computational experiments (cf. Section. 4).

Estimating the graph structure via reestimation and hard-thresholding: Although the L1 regularization (and constraint) set some of the coefficients to 0, they also retain a few entries with small absolute values. The following steps are used to improve the quality of the connectivity graph estimates returned by the LR and ISE procedures (Lokhov et al. 2018; Wu, Sanghavi, and Dimakis 2019).
(1) After selecting the edges connected to a node, the coefficients are reestimated by optimizing the unregularized (or unconstrained) objective only over these edges — that is by solving Problems (5) and (7) with $\lambda = 0$, and Problem (6) with $\lambda = \infty$.
(2) After deriving the symmetrized graph estimate, the authors only keep the entries with absolute weights larger than $\eta/2$, where η is the minimum absolute edge weight of $\boldsymbol{W}^*$:

$$\eta = \min\left\{|W_{ij}^*| : W_{ij}^* \neq 0\right\}. \tag{8}$$

Step (2) assumes the knowledge of η in the optimization procedure. The authors do not derive an estimator of this minimum weight, nor do they propose to select it using a validation set. The explicit use of η is then a shortcoming of all existing L1-based experiments. In contrast, we aim herein at deriving estimators of the connectivity graph which do not require the knowledge of η.

What this paper is about: Despite their good, well-studied, theoretical and empirical performance, both LR and ISE estimators rely on the use of a convex L1 regularization, and on the knowledge of η (cf. Equation (8)) to estimate sparse Ising graphs. In this work, we leverage insights from high-dimensional statistics (Rigollet 2015; Raskutti, Wainwright, and Yu 2011) and propose two novel estimators that learn the graph structure of sparse Ising models by solving L0-L2 constrained versions of Problems (5) and (7). The rest of this paper is organized as follows. In Section 2, we define our estimators as solutions of non-convex L0-L2 constrained problems and propose a discrete first order algorithm with low computational cost and convergence guarantees to obtain high-quality solutions. In Section 3, we prove that our estimators achieve the best upper bounds known for theoretical recovery of the connectivity graph of sparse Ising models: our bounds improve over existing rates for L1-based procedures. Finally, Section 4 assesses the computational performance of our estimators through a variety of experiments, for the graph topologies studied in the literature.

2 Cardinality Constrained Estimators

We consider the Ising model defined in Equation (1) where each node of the graph is of degree at most k^*. We aim at deriving sparse estimators of the connectivity matrix $\boldsymbol{W}^*$ with better statistical performance than L1-based LR and ISE. Our first estimator extends the LR procedure, and estimates the jth row of the connectivity matrix $\boldsymbol{w}_{-j}^*$ by solving the L0-L2 constrained logistic regression problem at node j:

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^{p-1}} \quad & \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-2y_i \mathbf{x}_i^T \boldsymbol{w})\right) \\ \text{s.t.} \quad & \|\boldsymbol{w}\|_0 \leq k, \; \|\boldsymbol{w}\|_2 \leq \lambda/\sqrt{k}, \end{aligned} \tag{9}$$

where $y_i = z_j^{(i)}$ and $\boldsymbol{x}_i = \boldsymbol{z}_{-j}^{(i)}$ as above. Problem (9) minimizes the logistic loss with a constraint on the number of non-zeros of $\boldsymbol{w}$ and a constraint on the L2 norm of $\boldsymbol{w}$. Similarly, our second estimator extends the ISE procedure, and solves the L0-L2 constrained ISE problem, defined for the jth node as

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^{p-1}} \quad & \frac{1}{n} \sum_{i=1}^{n} \exp(-y_i \mathbf{x}_i^T \boldsymbol{w}) \\ \text{s.t.} \quad & \|\boldsymbol{w}\|_0 \leq k, \; \|\boldsymbol{w}\|_2 \leq \lambda/\sqrt{k}. \end{aligned} \tag{10}$$

Problems (5), (6) and (7) consider one regularization parameter, λ, which simultaneously searches for a subset of features and shrinks the coefficients. In contrast, the L0 cardinality constraint in Problems (9) and (10) controls the model size while the L2 constraint controls the absolute values of the coefficients. The choice of the L2 constraint (over its L1 counterpart) is motivated by the theoretical guarantees of our estimators, which we present in Section 3.

Solving L0-constrained problems: The use of the L0 constraint has been widely studied for the least-squares settings

(Friedman, Hastie, and Tibshirani 2001). Despite its excellent statistical properties (Bunea et al. 2007; Raskutti, Wainwright, and Yu 2011), L0-constrained least-squares (aka *best subsets*) has often been perceived as computationally infeasible (Natarajan 1995). Consequently, L1 regularization (aka *Lasso*) (Tibshirani 1996) has often been proposed as a convex surrogate to promote sparsity, while achieving near-optimal statistical performance (Bickel et al. 2009). However, recent work has shown that mixed integer programming algorithms can solve L0-constrained regression and classification problems with $p \sim 10^6$ variables (Hazimeh and Mazumder 2018; Bertsimas, Pauphilet, and Van Parys 2017; Dedieu, Hazimeh, and Mazumder 2020) in times comparable to fast L1-based algorithms, while leading to large performance improvements. In addition, pairing L0 with an additional L1 or L2 penalty has been proved to drastically improve the performance of the standalone best subsets (Mazumder, Radchenko, and Dedieu 2017). We leverage these results in designing our proposed estimators.

Despite a rich body of work for least-squares, the L0 penalty has never been used for recovering the graph structure of Ising models. We aim herein at bridging this gap by proposing the connectivity matrix estimates defined as solutions of Problems (9) and (10). In particular, we derive below a discrete first order algorithm that extends an existing framework and leverages the use of warm-starts to compute high-quality solutions for these non-convex problems.

2.1 Discrete First Order Algorithm

Inspired by proximal gradient methods for convex optimization (Nesterov 2013), we propose a discrete first order (DFO) algorithm for obtaining good estimates of $\boldsymbol{W}^*$ by solving the L0-L2 constrained Problems (9) and (10). Our procedure adapts an existing framework (Bertsimas, King, and Mazumder 2016), has low computational complexity and uses warm-start to increase the quality of the solutions.

We assume that f is a non-negative convex differentiable loss with C-Lipschitz continuous gradient, that is, it satisfies:

$$\|\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{v})\|_2 \leq C \|\boldsymbol{w} - \boldsymbol{v}\|_2 \quad \forall \boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^{p-1}. \quad (11)$$

When $f(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-2y_i \mathbf{x}_i^T \boldsymbol{w})\right)$ is the logistic loss, the Hessian of f can be expressed as:

$$\nabla^2 f(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \frac{4 \exp(-2y_i \mathbf{x}_i^T \boldsymbol{w})}{\left(1 + \exp(-2y_i \mathbf{x}_i^T \boldsymbol{w})\right)^2} \mathbf{x}_i \mathbf{x}_i^T,$$

and we can use $C = n^{-1} \sigma_{\max}(\boldsymbol{X}^T \boldsymbol{X})$, where we have noted $\boldsymbol{X} \in \mathbb{R}^{n \times (p-1)}$ the matrix with ith row $\boldsymbol{x}_i$, and where $\sigma_{\max}(.)$ is the maximum eigeinvalue of a matrix. Similarly, the Hessian of the interaction screening loss is:

$$\nabla^2 f(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \exp(-y_i \mathbf{x}_i^T \boldsymbol{w}) \mathbf{x}_i \mathbf{x}_i^T,$$

for which we can use $C = n^{-1} e^{\lambda} \sigma_{\max}(\boldsymbol{X}^T \boldsymbol{X})$. We describe a DFO method for the following problem:

$$\min f(\boldsymbol{w}) \quad \text{s.t.} \quad \|\boldsymbol{w}\|_0 \leq k, \ \|\boldsymbol{w}\|_2 \leq \theta, \quad (12)$$

where $\theta \geq 0$. For $D \geq C$, we upper-bound the smooth f around any $\boldsymbol{v} \in \mathbb{R}^{p-1}$ with the quadratic form $Q_D(\boldsymbol{v}, .)$ defined $\forall \boldsymbol{w} \in \mathbb{R}^{p-1}$ as

$$f(\boldsymbol{w}) \leq Q_D(\boldsymbol{v}, \boldsymbol{w}) = f(\boldsymbol{v}) + \nabla f(\boldsymbol{v})^T (\boldsymbol{w} - \boldsymbol{v}) + \frac{D}{2} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2.$$

Given a solution $\boldsymbol{v}$, our method minimizes the upper-bound of f around $\boldsymbol{v}$. That is, it solves:

$$\begin{aligned} \hat{\boldsymbol{w}} &\in \underset{\boldsymbol{w}: \|\boldsymbol{w}\|_0 \leq k, \|\boldsymbol{w}\|_2 \leq \theta}{\operatorname{argmin}} Q_D(\boldsymbol{v}, \boldsymbol{w}) \\ &\in \underset{\boldsymbol{w}: \|\boldsymbol{w}\|_0 \leq k, \|\boldsymbol{w}\|_2 \leq \theta}{\operatorname{argmin}} \left\| \boldsymbol{w} - \left(\boldsymbol{v} - \frac{1}{D} \nabla f(\boldsymbol{v})\right) \right\|_2^2. \end{aligned} \quad (13)$$

The above can be solved with the the proximal operator:

$$\mathcal{S}(\boldsymbol{v}; k; \theta) := \underset{\boldsymbol{w}: \|\boldsymbol{w}\|_0 \leq k, \|\boldsymbol{w}\|_2 \leq \theta}{\operatorname{argmin}} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2, \quad (14)$$

which can be computed with Proposition 1. The proof is presented in the Supplementary Materials. As the ordering of $|v_j|$ may have ties, the solution of Problem (13) may be non-unique, and the solution of Problem (14) is set-based.

Proposition 1 *Let* $(1), \ldots, (p)$ *be a permutation of the indices* $1, \ldots, p$ *such that the entries in* $\boldsymbol{v}$ *are sorted as:* $|v_{(1)}| \geq |v_{(2)}| \geq \ldots \geq |v_{(p)}|$. *Then,* $\hat{\boldsymbol{w}} \in \mathcal{S}(\boldsymbol{v}; k; \theta)$ *is given by:*

$$\hat{w}_i = \begin{cases} \min\left(1, \frac{\theta}{\tau_v}\right) v_i & i \in \{(1), (2), \ldots, (k)\} \\ 0 & otherwise, \end{cases}$$

where $\tau_v = \sqrt{\sum_{i=1}^{k} v_{(i)}^2}$ *is the L2 norm of the* k *largest (absolute) entries of* $\boldsymbol{v}$.

DFO algorithm: The DFO algorithm starts from an initialization $\boldsymbol{w}^{(1)}$ and performs the following updates (for $t \geq 1$)

$$\boldsymbol{w}^{(t+1)} \in \mathcal{S}\left(\boldsymbol{w}^{(t)} - \frac{1}{D} \nabla f(\boldsymbol{w}^{(t)}); k; \theta\right),$$

until some convergence criterion is met. In practice, we stop the algorithm when $\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\|_2^2 \leq \epsilon$ for some small ϵ or when we have reached a maximum number $T_{\max}$ of iterations. Let $\tilde{\boldsymbol{w}}$ denote the estimator returned and $\mathcal{I}(\tilde{\boldsymbol{w}}) = \{i : \tilde{w}_i \neq 0\}$ be the set of edges selected — which is of size at most k. The last step of the DFO algorithm reestimates the weights estimates of $\mathcal{I}(\tilde{\boldsymbol{w}})$ by solving the convex problem

$$\min f(\boldsymbol{w}) \quad \text{s.t.} \quad w_i = 0, \ \forall i \notin \mathcal{I}(\tilde{\boldsymbol{w}}).$$

2.2 Convergence Properties

We establish convergence properties of the sequence $\{\boldsymbol{w}^{(t)}\}$ in terms of reaching a first order stationary point. Our work adapts the existing framework from Bertsimas, King, and Mazumder (2016) to the constrained Problem (12). We first consider the following definition.

Definition 1 *We say that* $\boldsymbol{w}$ *is a first order stationary point of Problem* (12) *if* $\boldsymbol{w} \in \mathcal{S}(\boldsymbol{w} - \frac{1}{L} \nabla f(\boldsymbol{w}); k; \theta)$. *We say that* $\boldsymbol{w}$ *is an* ϵ-*accurate first order stationary point if* $\|\boldsymbol{w}\|_0 \leq k$; $\|\boldsymbol{w}\|_2 \leq \theta$, *and* $\left\|\boldsymbol{w} - \mathcal{S}\left(\boldsymbol{w} - \frac{1}{D} \nabla f(\boldsymbol{w}); k; \theta\right)\right\|_2^2 \leq \epsilon$.

We can now derive the following proposition. The proof is presented in the Supplementary Materials.

Proposition 2 *Let $\{\boldsymbol{w}^{(t)}\}_{1 \leq t \leq T}$ be a sequence generated by the DFO algorithm. Then, for $D > C$, the sequence $f(\boldsymbol{w}^{(t)})$ is decreasing, and it converges to some $f^* \geq 0$. In addition, the following convergence rate holds:*

$$\min_{1 \leq t \leq T-1} \|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\|_2^2 \leq \frac{2(f(\boldsymbol{w}^{(1)}) - f^*)}{T(D - C)}.$$

Consequently, the algorithm reaches an ϵ-accurate first order stationary point (cf. Definition 1) in $O(\epsilon^{-1})$ iterations. In practice, the DFO algorithm converges much faster than the sublinear rate suggested by Proposition 2. This is the case when we leverage the use of warm-starts.

2.3 Continuation Heuristic via Warm-starts

Because Problem (12) is non-convex, the DFO algorithm is sensitive to the initialization $\boldsymbol{w}^{(1)}$. We design herein a continuation scheme which both (a) improves the quality of the solution and (b) decreases the computational cost when solving Problem (12) at a particular value of k. We consider a decreasing list of parameters $\{k_1, \ldots, k_r\}$ with $k_1 = p$ and $k_r = 1$. For $k_1 = p$, we solve Problem (12) without the cardinality constraint[2]. Our method iteratively refines the solution $\hat{\boldsymbol{w}}(p)$ and returns a sequence of estimates $\{\hat{\boldsymbol{w}}(k_i)\}_{1 \leq i \leq r}$ with decreasing support sizes. In particular, Problem (12) at k_r is solved by warm-starting with the solution obtained from k_{r-1}. We summarize the approach below.

Continuation heuristic
(1) Initialize $\hat{\boldsymbol{w}}(k_1)$ by solving Problem (12) without the cardinality constraint.
(2) For $i = 2, \ldots r$, set $\theta = 2\|\hat{\boldsymbol{w}}(k_{i-1})\|_1$.
Set $\hat{\boldsymbol{w}}(k_i)$ as the output of the DFO algorithm initialized with $\hat{\boldsymbol{w}}(k_{i-1})$, for k_i and the above value of θ.

Specification to LR and ISE: The above heuristic can be applied to Problems (9) and (10) to decrease the runtime of our DFO algorithm while returning higher-quality solutions. In practice we set $\hat{\boldsymbol{w}}(p)$ as the respective solution of Problem (5) for LR and Problem (7) for ISE. In addition, the value of k^* is unknown. As we return a sequence of estimators with decreasing support sizes, we propose to derive an estimator $\hat{k}$ of k^* by minimizing the Bayesian information criterion (Schwarz et al. 1978). Our estimate $\hat{\boldsymbol{w}}(\hat{k})$ will naturally be $\hat{k}$ sparse. Hence, contrary to all the L1-based approaches, we do not need the knowledge of η (cf. Equation (8)) to estimate the connectivity graph (cf. Section 4).

3 Statistical Properties

In this section, we study the statistical performance of the proposed L0-L2 constrained estimators for recovering the connectivity graph of the Ising model defined in Equation (1). In particular, we propose an upper bound on the number of samples required for high probability recovery. Our results do not require any external assumption and leverage the excellent statistical properties of the L0 constraint

---

[2]The value of θ can be tuned on an independent validation set.

— which are well-known for the sparse least-squares settings (Bunea et al. 2007; Raskutti, Wainwright, and Yu 2011; Dedieu, Hazimeh, and Mazumder 2020) — to improve the best existing bounds, achieved by L1-based estimators.

Notations: We use the minimum edge weight η defined in Equation (8). We additionally define the sparsity and width (k^*, λ^*) of the connectivity matrix $\boldsymbol{W}^*$ as

$$\inf\left\{(k, \lambda) \,\middle|\, \|\boldsymbol{w}^*_{-j}\|_0 \leq k, \ \sqrt{k}\|\boldsymbol{w}^*_{-j}\|_2 \leq \lambda, \ \forall j\right\}$$

where we have used the lexicographic ordering. In particular, it holds $\|\boldsymbol{w}^*_{-j}\|_1 \leq \lambda^*, \forall j$.

3.1 Comparisons with Existing Work

Table 1 compares the sample complexity of existing approaches for recovering sparse Ising graphs with high probability. Because k^* and λ^* are unknown, all methods consider a couple (k, λ) where $k \geq k^*$, and $\lambda \geq \lambda^*$. The table does not mention the L1-regularized LR estimator (Ravikumar et al. 2010) defined as a solution of Problem (5) as it relies on incoherence assumptions, which are NP-hard to validate (Bandeira et al. 2013). The L1-constrained LR (Rigollet 2015; Wu, Sanghavi, and Dimakis 2019) defined as a solution of Problem (6) has been proved to reach the best upper bound known for sample complexity, and to be the most sample-efficient approach. Its bound scales exponentially with the width λ, logarithmically with the number of

Estimator + paper	Sample complexity
Greedy method (Bresler 2015) $\star$	$O\left(\exp\left(\frac{\exp(O(k\lambda))}{\eta^{O(1)}}\right)\log\left(\frac{p}{\delta}\right)\right)$
L1 ISE (Vuffray et al. 2016) $\star$	$O\left(\max\left(k, \frac{1}{\eta^2}\right)k^3 e^{6\lambda}\log\left(\frac{p}{\delta}\right)\right)$
L1 LR (Lokhov et al. 2018) $\star$	$O\left(\max\left(k, \frac{1}{\eta^2}\right)k^3 e^{8\lambda}\log\left(\frac{p}{\delta}\right)\right)$
L1-constrained LR (Rigollet 2015)	$O\left(\frac{\lambda^2 \exp(8\lambda)}{\eta^4}\log\left(\frac{p}{\delta}\right)\right)$
L1-constrained LR (Wu, Sanghavi, and Dimakis 2019)	$O\left(\frac{\lambda^2 \exp(12\lambda)}{\eta^4}\log\left(\frac{p}{\delta}\right)\right)$
L0-L2 constrained LR (this paper) $\star$	$O\left(\frac{\lambda^2 \exp(8\lambda)}{\eta^4}\log\left(\frac{p}{\sqrt{k}}\right)\log\left(\frac{2}{\delta}\right)\right)$
L0-L2 constrained ISE (this paper) $\star$	$O\left(\frac{\lambda^2 \vee \lambda^4 \exp(8\lambda)}{\eta^4}\log\left(\frac{p}{\sqrt{k}}\right)\log\left(\frac{2}{\delta}\right)\right)$

Table 1: Comparison of the sample complexity required for graph recovery guarantees with probability at least $1 - \delta$, for an Ising model with p nodes, width λ^* and minimum absolute edge weight η. Part of the list is adapted from Wu, Sanghavi, and Dimakis (2019). Papers referred with a $\star$ additionally assume the degree of each node to be bounded by k^*. All approaches consider $\lambda \geq \lambda^*$ and $k \geq k^*$.

nodes p and shows a polynomial dependency on $1/\eta$. In this section, we prove that our proposed L0-L2 constrained estimators need fewer samples than all the existing approaches to recover the connectivity matrix with high probability. Our results are summarized in the last two rows of Table 1.

Connection with least-squares: The intuition behind the improved statistical performance of L0-L2 LR and L0-L2 ISE for recovering sparse Ising models can be derived from the least-squares problem. In this case, we consider the task of recovering a k sparse vector in $\mathbb{R}^p$ from n independent realizations of a Gaussian linear model. For mean squared error, best subsets achieves a rate of $O\left((k/n)\log(p/k)\right)$ which is known to be optimal (Ravikumar et al. 2010; Rigollet 2015). In contrast, Lasso achieves a rate of $O\left((k/n)\log(p)\right)$ under restricted eigenvalue assumptions (Bickel et al. 2009), which are NP-hard to satisfy.

3.2 Upper Bound for L0-L2 Constrained LR

The following theorem summarizes our main result for learning the connectivity matrix of an Ising model with the L0-L2 LR procedure. We refer to the Supplementary Materials for a detailed presentation and proof.

Theorem 1 *Let $\delta \in (0, 1)$. The L0-L2 constrained logistic regression estimator of the connectivity matrix $\hat{W}_{\mathrm{LR}}$, defined by solving Problem* (9) *for all nodes for the parameters $k \geq k^*, \lambda \geq \lambda^*$, satisfies with probability at least $1 - \delta$:*

$$\|\hat{W}_{\mathrm{LR}} - W^*\|_\infty^2 \lesssim \lambda e^{4\lambda} \sqrt{\frac{\log(p/\sqrt{k})}{n}\log(2/\delta)}.$$

The constant in the bound is defined in the appendices. We derive our upper bound on the number of samples required to recover with high probability the connectivity graph.

Corollary 1 *Let $\delta \in (0, 1)$. Assume that the number of samples satisfies $n = O\left(\frac{\lambda^2 \exp(8\lambda)}{\eta^4}\log\left(\frac{p}{\sqrt{k}}\right)\log\left(\frac{2}{\delta}\right)\right)$. The L0-L2 constrained logistic regression estimator derived by hard-thresholding $\hat{W}_{\mathrm{LR}}$ for the threshold $\eta/2$ recovers the exact connectivity graph with probability at least $1 - \delta$.*

When $k = k^*$ in Problem (9), the L0-L2 LR estimator does not need the knowledge of η to estimate the connectivity graph. In practice, we select k to minimize some information criterion and do not use the thresholding step (cf. Section 4).

3.3 Upper Bound for L0-L2 Constrained ISE

Theorem 2 derives a similar upper bound than above when estimating the connectivity matrix of the Ising model in Equation (1) with the L0-L2 ISE procedure.

Theorem 2 *Let $\delta \in (0, 1)$. The L0-L2 constrained interaction screening estimator of the connectivity matrix $\hat{W}_{\mathrm{ISE}}$, defined by solving Problem* (10) *for all nodes for the parameters $k \geq k^*, \lambda \geq \lambda^*$ satisfies with probability at least $1 - \delta$:*

$$\|\hat{W}_{\mathrm{ISE}} - W^*\|_\infty^2 \lesssim (\lambda \vee \lambda^2) e^{4\lambda} \sqrt{\frac{\log(p/\sqrt{k})}{n}\log(2/\delta)}.$$

The sample complexity of the L0-L2 ISE follows.

Corollary 2 *Let $\delta \in (0, 1)$. Assume that the number of samples satisfies $n = O\left(\frac{(\lambda^2 \vee \lambda^4)\exp(8\lambda)}{\eta^4}\log\left(\frac{p}{\sqrt{k}}\right)\log\left(\frac{2}{\delta}\right)\right)$. The L0-L2 constrained interaction screening estimator derived by thresholding $\hat{W}_{\mathrm{ISE}}$ for the threshold $\eta/2$ recovers the exact connectivity graph with probability at least $1 - \delta$.*

As presented in Table 1, both our proposed estimators achieve better upper bounds and are more sample-efficient than the existing approaches studied in the literature for learning Ising models. In particular, the dependency on the dimension size is lowered from a $\log(p)$ scaling to $\log(p/\sqrt{k})$. This is explained by the fact that the number of samples required for each node is lowered from a $\log(p)$ dependency to a $\log(p/k)$ one. The upper bounds then hold by applying a union bound on the p nodes of the graph.

Remark: The decreasing dependency of our bounds with k comes from the decreasing dependency of the L2 constraint with k in Problems (9) and (10). This allows us to have a model width bounded by λ, as in the literature. We could alternatively use a fixed L2 constraint with value λ, and replace the model width bound by $\lambda\sqrt{k}$ for each method in Table 1. Our bounds would now increase with k, while still improving over the existing ones for L1-based procedures.

4 Computational Experiments

In this section, we assess the empirical performance of the proposed estimators for recovering the underlying graphs of various classes of Ising models. All the experiments were run on an Amazon Web Service c5.9 instance with 3.6GHz Xeon Platinum 8000 processor, 72GB of RAM.

4.1 Data Generation

We use the graph topologies that have been studied in the literature (Vuffray et al. 2016; Lokhov et al. 2018) and consider the two following classes of connectivity graphs G:

Figure 1: Graph topologies for Examples 1 and 2

Example 1 We assume that p is the square of an integer, and consider a four-connected two-dimensional lattice of size $\sqrt{p}$ with periodic boundary conditions. Each edge is of degree $k^* = 4$. All the couplings take a similar value $\eta = 0.5^3$.

Example 2: We consider a random regular graph with degree $k^* = 3$. Couplings take random values that are uniformly generated in the range $[0.7, 0.9]$.

---

[3]This setting has been proposed to compare LR with ISE (Lokhov et al. 2018). We decrease the value of η from 0.7 to 0.5 to decrease the computational cost as (a) it does not affect the relative performance of the methods and (b) our experiments are more consuming than the authors as we tune the parameters.
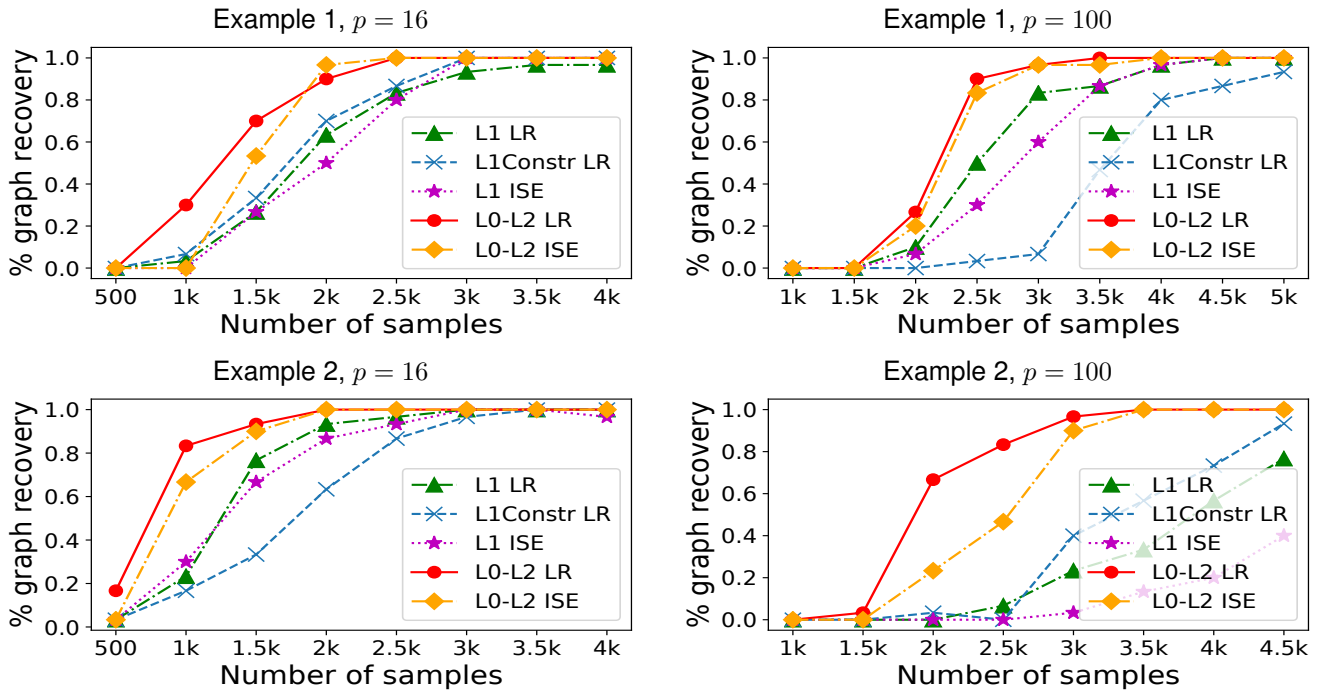
Figure 2: Examples 1 and 2 with a small and a large number of nodes. The ratio of success is estimated over 30 repetitions. Our proposed L0-L2 constrained estimators do not use the value of the minimum absolute edge weight η to estimate the graph structure. Both outperform all their L1-based counterparts for both topologies while exhibiting sharper phase transitions.

We simulate n independent realizations from an Ising model with the above connectivity matrices. When $p \leq 16$, the partition function defined in Equation (2) is computationally tractable and the samples are exactly drawn from the Ising model. When $p > 16$, each observation is generated by running 1000 iterations of a Gibbs sampler, exploiting the dynamics described in Equation (3).

4.2 Methods Compared

We compare the following methods in our experiments:

• **L1 LR:** This is the L1-regularized logistic regression estimator (Ravikumar et al. 2010) defined for each node as a solution of Problem (5), which we solve using Python's `scikit-learn` package (Pedregosa et al. 2011). For each node, we select the regularization parameter leading to the highest conditional likelihood on an independent validation set — rather than setting it to the value suggested in Lokhov et al. (2018) — as we observed that the former leads to better performance. More precisely, we compute a family of estimators for a decreasing geometric sequence of 20 parameters $\lambda_1, \ldots, \lambda_{20}$ with common ratio 0.5. We start from $\lambda_1 = 2\|\boldsymbol{X}\boldsymbol{y}\|_\infty$ for which the solution of Problem (5) is $\boldsymbol{0}$.

• **L1Constr LR:** This is the L1-constrained logistic regression estimator (Rigollet 2015; Wu, Sanghavi, and Dimakis 2019) defined for each node as a solution of Problem (6). We implement a first order algorithm, using the FISTA acceleration procedure (Beck and Teboulle 2009) for faster convergence. We use a stopping criterion $\epsilon = 10^{-3}$

and a maximum number of $T_{\max} = 300$ iterations. Projection onto the L1 ball is achieved by the SPGL1 algorithm (van den Berg and Friedlander 2008) using the software provided (van den Berg and Friedlander 2019). We additionally tune the constraint parameter over the validation set.

• **L1 ISE:** This is the L1-regularized interaction screening estimator (Vuffray et al. 2016; Lokhov et al. 2018) defined for each node as a solution of Problem (7). Similarly to the above, we use an accelerated first-order algorithm and select, for each node, the regularization parameter leading to the highest conditional likelihood on the same validation set — as we observed better empirical performance than using the value suggested in Lokhov et al. (2018).

• **L0-L2 LR:** This is the proposed L0-L2 constrained logistic regression estimator defined for each node as a solution of Problem (9). We pair the DFO algorithm presented in Section 2.1 — with a stopping criterion $\epsilon = 10^{-3}$ and a maximum number of $T_{\max} = 300$ iterations — with the continuation heuristic described in Section 2.3. For $k = p$, we initialize the heuristic with the L1 LR estimator. The value of k is selected by minimizing the Bayesian information criterion (BIC) (Schwarz et al. 1978) defined as $\text{BIC}(k) = \log(n)S(k) - 2\log(\mathcal{L})$ where $\mathcal{L}$ is the train pseudo-likelihood, n the training set size and $S(k)$ the total number of edges of the estimated connectivity graph when the constraint is that each node is of degree at most k.

• **L0-L2 ISE:** This is the proposed L0-L2 constrained interaction screening estimator, defined for each node as a
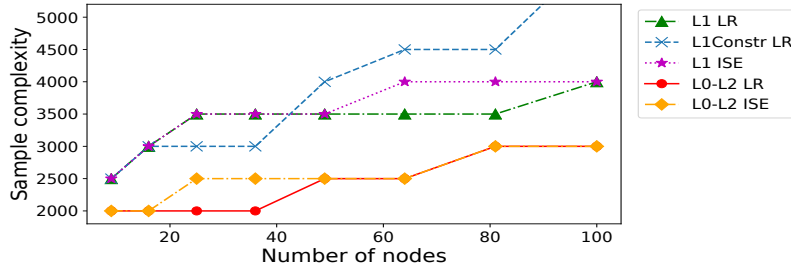
Figure 3: Example 1 with varying graph sizes. Both our L0-L2 constrained estimators constantly need fewer samples than their L1-based counterparts to fully recover the Ising graph structures.

solution of Problem (10). The computational and tuning procedures are similar to the ones for L0-L2 LR. We initialize the continuation heuristic for $k = p$ with L1 ISE.

Estimating the graph structure: As presented in Section 1, after solving each problem with the use of the corresponding regularization or constraint, we reestimate the objective without penalty only over the selected edges, and derive a symmetrized estimate by sharing the parameters for both halves. More importantly, because L0-L2 LR and L0-L2 ISE are naturally sparse, they do not need the knowledge of the minimum edge weight η (cf. Equation (8)). In contrast, for all the L1-based estimators, as described in their respective papers, we form the connectivity graph estimate by keeping the edges with absolute weights larger than $\eta/2$.

4.3 Phase Transition as n Increases

Our first experiment compares the five different methods for Examples 1 and 2 and an increasing number of samples. More precisely, for a fixed value of p, the values of n considered follow an increasing arithmetic sequence with common difference 500. For each value of n, we simulate 30 independent[4] connectivity matrices $\boldsymbol{W}^*$ and training sets of size n. We additionally generate a validation set of the same size than the training set, using the same $\boldsymbol{W}^*$. For each run and each method, we select the best estimator as explained above and report the binary observation of whether the graph structure is fully recovered. We then compute the ratio of success for each method and each value of n over the 30 realizations.

We report our findings in Figure 2. We consider Examples 1 and 2 for a small number of nodes $p = 16$ and a large number of nodes $p = 100$. All the methods in Figure 2 show a phase transition: their performance increases as the number of samples increases. In addition, for both examples and both sizes, our two proposed L0-L2 constrained estimators show better empirical performance and sharper phase transitions than all the L1-based procedures. In particular, our methods both need fewer samples than their counterparts to exactly recover the Ising graph structures.

We complement our findings by reporting the L2 norm of the difference between $\boldsymbol{W}^*$ and the connectivity matrix

estimate of each method (before thresholding for L1-based ones) in the Supplementary Materials. While this metric — referred to as L2 estimation — does not need the knowledge of η (for L1-based procedures), our estimators still show important gains in performance.

4.4 Performance Gain for All Graph Sizes

The above suggests that, for $p \in \{16, 100\}$, our L0-L2 constrained estimators need fewer samples than L1-based methods to recover sparse Ising graphs. We explore herein whether this property holds for all graph sizes. For an estimator and a graph size p, the sample complexity $m^*(p)$ has been defined (Vuffray et al. 2016; Lokhov et al. 2018) as the smallest number of samples the estimator requires to learn the connectivity graph of a model with p nodes for all runs. We propose to use the more robust metric where we allow at most three failures (e.g. 90% success) over the 30 repetitions. We additionally define the sample complexity over all graph sizes smaller than p as $n^*(p) = \max_{q \leq p} m^*(q)$.

We consider Example 1 with the sequence of p values $\{9, 16, 25, 36, 49, 64, 81, 100\}$ and run the same experiments as in Section 4.3. For each method and each value of p, we compute the corresponding value of $n^*(p)$ defined above. We report our findings in Figure 3. For every graph size, our proposed L0-L2 constrained estimators both outperform all the L1-based procedures. That is, our approaches need fewer samples for recovering the Ising graphs studied in Vuffray et al. (2016); Lokhov et al. (2018) than the best known methods. In addition, as suggested by our theoretical results, each estimator needs more samples to recover larger graphs. We also note that (a) L1Constr LR performance decrease when p increases and (b) the validation procedure does not give a winner between L1 LR and L1 ISE contrary to what has been reported with a fixed theoretically-driven parameter. Finally, we show in the Supplementary Materials that, for every value of p, our estimators achieve sharper phase transitions and important gains for L2 estimation.

5 Further Work

Although this paper focuses on learning binary Ising models, an interesting follow-up work would extend the proposed estimators for learning general discrete graphical models (Ravikumar et al. 2010; Wu, Sanghavi, and Dimakis 2019).

---

[4]Note that from the rule of three in statistics, the absence of failure of an event over 30 repetitions guarantees a 95% confidence interval where the probability of success is larger than 0.9.

Acknowledgements

References

Bandeira, A. S.; Dobriban, E.; Mixon, D. G.; and Sawin, W. F. 2013. Certifying the restricted isometry property is hard. *IEEE transactions on information theory* 59(6): 3448–3450.

Beck, A.; and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2(1): 183–202.

Bertsimas, D.; King, A.; and Mazumder, R. 2016. Best subset selection via a modern optimization lens. *The annals of statistics* 813–852.

Bertsimas, D.; Pauphilet, J.; and Van Parys, B. 2017. Sparse classification and phase transitions: A discrete optimization perspective. *arXiv preprint arXiv:1710.01352* .

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

Bickel, P. J.; Ritov, Y.; Tsybakov, A. B.; et al. 2009. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of statistics* 37(4): 1705–1732.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

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

Bunea, F.; Tsybakov, A. B.; Wegkamp, M. H.; et al. 2007. Aggregation for Gaussian regression. *The Annals of Statistics* 35(4): 1674–1697.

Dedieu, A.; Hazimeh, H.; and Mazumder, R. 2020. Learning sparse classifiers: Continuous and mixed integer optimization perspectives. *arXiv preprint arXiv:2001.06471* .

Friedman, J.; Hastie, T.; and Tibshirani, R. 2001. *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Hazimeh, H.; and Mazumder, R. 2018. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *arXiv preprint arXiv:1803.01454* .

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

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

Mazumder, R.; Radchenko, P.; and Dedieu, A. 2017. Subset selection with shrinkage: Sparse linear modeling when the SNR is low. *arXiv preprint arXiv:1708.03288* .

McCoy, B. M.; and Maillard, J.-M. 2012. The importance of the Ising model. *Progress of Theoretical Physics* 127(5): 791–817.

Montanari, A.; and Pereira, J. A. 2009. Which graphical models are difficult to learn? In *Advances in Neural Information Processing Systems*, 1303–1311.

Natarajan, B. K. 1995. Sparse approximate solutions to linear systems. *SIAM journal on computing* 24(2): 227–234.

Nesterov, Y. 2013. Gradient methods for minimizing composite functions. *Mathematical Programming* 140(1): 125–161.

Onsager, L. 1944. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review* 65(3-4): 117.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Raskutti, G.; Wainwright, M. J.; and Yu, B. 2011. Minimax rates of estimation for high-dimensional linear regression over lq-balls. *IEEE transactions on information theory* 57(10): 6976–6994.

Ravikumar, P.; Wainwright, M. J.; Lafferty, J. D.; et al. 2010. High-dimensional Ising model selection using L1-regularized logistic regression. *The Annals of Statistics* 38(3): 1287–1319.

Rigollet, P. 2015. 18.S997: High Dimensional Statistics. *Lecture Notes), Cambridge, MA, USA: MIT OpenCourseWare* .

Schwarz, G.; et al. 1978. Estimating the dimension of a model. *The annals of statistics* 6(2): 461–464.

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

Thompson, C. J. 2015. *Mathematical statistical mechanics*. Princeton University Press.

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

van den Berg, E.; and Friedlander, M. P. 2008. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing* 31(2): 890–912. doi:10.1137/080714488. URL http://link.aip.org/link/?SCE/31/890.

van den Berg, E.; and Friedlander, M. P. 2019. SPGL1: A solver for large-scale sparse reconstruction. Https://friedlander.io/spgl1, visited 2020-07-14.

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*, 2595–2603.

Welling, M.; and Sutton, C. A. 2005. Learning in Markov Random Fields with Contrastive Free Energies. In *AISTATS*.

Wu, S.; Sanghavi, S.; and Dimakis, A. G. 2019. Sparse logistic regression learns all discrete pairwise graphical models. In *Advances in Neural Information Processing Systems*, 8071–8081.