arXiv:1810.11905v1 [cs.LG] 28 Oct 2018

Sparse Logistic Regression Learns All Discrete Pairwise Graphical Models

Shanshan Wu, Sujay Sanghavi, Alexandros G. Dimakis
shanshan@utexas.edu, sanghavi@mail.utexas.edu,
dimakis@austin.utexas.edu

Department of Electrical and Computer Engineering
University of Texas at Austin

Abstract

We characterize the effectiveness of a natural and classic algorithm for recovering the Markov graph of a general discrete pairwise graphical model from i.i.d. samples. The algorithm is (appropriately regularized) conditional maximum likelihood, which involves solving a convex program for each node; for Ising models this is ℓ_1-constrained logistic regression, while for more alphabets an $\ell_{2,1}$ group-norm constraint needs to be used.

We show that this algorithm can recover any arbitrary discrete pairwise graphical model, and also characterize its sample complexity as a function of model width, alphabet size, edge parameter accuracy, and the number of variables. We show that along every one of these axes, it matches or improves on all existing results and algorithms for this problem. Our analysis applies a sharp generalization error bound for logistic regression when the weight vector has an ℓ_1 constraint (or $\ell_{2,1}$ constraint) and the sample vector has an ℓ_∞ constraint (or $\ell_{2,\infty}$ constraint). We also show that the proposed convex programs can be efficiently optimized in $\tilde{O}(n^2)$ running time (where n is the number of variables) under the same statistical guarantee. Our experimental results verify our analysis.

1 Introduction

An undirected graphical model, or Markov random field (MRF), provides a general framework for modeling the interaction between random variables. It has applications in a wide range of areas, including computer vision [CLTW10], bio-informatics [MCK+12], and sociology [EPL09].

In this paper we characterize the effectiveness of a natural, and already popular, algorithm for the *structure learning* problem, in general discrete pairwise graphical models. These are MRFs where the variables can take values in general discrete alphabets, but all interactions are pairwise. This includes the Ising model as a special case. Structure learning is the task of finding the underlying dependency graph of an MRF given i.i.d. samples; typically one is also interested in finding estimates for the edge parameters as well.

The natural and popular algorithm we consider is (appropriately regularized) conditional maximum likelihood for finding the neighborhood set of any given node. For the Ising model, this becomes ℓ_1-constrained logistic regression; more generally for non-binary models the regularizer has to be an $\ell_{2,1}$ norm. We show that this algorithm can recover all discrete pairwise graphical models, and characterize its sample complexity as a function of the parameters of interest: model width, alphabet size, edge parameter accuracy, and the number of variables. We match or improve dependence on each of these parameters, over all existing results for the general pairwise discrete case when no additional assumptions are made on the model. For the specific case of Ising models, some recent work has better dependence on some parameters (see Appendix A).

We now describe the related work, and then outline our contributions.

Related Work

In a classic paper, Ravikumar, Wainwright and Lafferty [RWL10] considered the structure learning problem for Ising models. They showed that ℓ_1-regularized logistic regression provably recovers the correct dependency graph with a very small number of samples by solving a convex program for each variable. This algorithm was later generalized to multi-class logistic regression with group-sparse regularization, which can learn MRFs with higher-order interactions and non-binary variables [JRVS11]. A well-known limitation of [RWL10, JRVS11] is that the theoretical guarantees only work for a restricted class of graphs. Specifically, they require that the underlying graph satisfies technical *incoherence* assumptions, that are difficult to validate or check.

Paper	Assumptions	Sample complexity (N)
Greedy algorithm [HKM17]	1. Alphabet size $k \geq 2$ 2. Model width $\leq \lambda$ 3. Degree $\leq d$ 4. Minimum edge weight $\geq \eta > 0$ 5. Probability of success $\geq 1 - \rho$	$O(\exp(\frac{k^{O(d)} \exp(O(d^2\lambda))}{\eta^{O(1)}}) \ln(\frac{nk}{\rho}))$
Sparsitron [KM17]	1. Alphabet size $k \geq 2$ 2. Model width $\leq \lambda$ 3. Minimum edge weight $\geq \eta > 0$ 4. Probability of success $\geq 1 - \rho$	$O(\frac{\lambda^2 k^5 \exp(14\lambda)}{\eta^4} \ln(\frac{nk}{\rho\eta}))$
$\ell_{2,1}$-constrained logistic regression [this paper]	1. Alphabet size $k \geq 2$ 2. Model width $\leq \lambda$ 3. Minimum edge weight $\geq \eta > 0$ 4. Probability of success $\geq 1 - \rho$	$O(\frac{\lambda^2 k^4 \exp(14\lambda)}{\eta^4} \ln(\frac{nk}{\rho}))$

Table 1: Comparison of sample complexity for graph recovery of a discrete pairwise graphical model with alphabet size k. For $k = 2$ (i.e., Ising models), our algorithm reduces to the ℓ_1-constrained logistic regression (see Appendix A for related work on the special case of learning Ising models). Our sample complexity has a better dependency on the alphabet size ($\tilde{O}(k^4)$ versus $\tilde{O}(k^5)$) than that in [KM17].

A large amount of recent work has since proposed various algorithms to obtain provable learning results for general graphs without requiring incoherence assumptions. We now describe the (most related part of the extensive) related work, followed by our results and comparisons (see Table 1). For a discrete pairwise graphical model, let n be the number of variables and

k be the alphabet size; define the model width λ as the maximum neighborhood weight (see Definition 1 and 2 for the precise definition). For the case of $k = 2$ (i.e., Ising models), Santhanam and Wainwright [SW12] provided an information-theoretic lower bound on the number of samples, which depends logarithmically on n, and exponentially on the width λ.

For structure learning algorithms, a popular approach is to focus on the sub-problem of finding the neighborhood of a single node. Once this is correctly learned, the overall graph structure is a simple union bound. Indeed all the papers we now discuss are of this type. As shown in Table 1, Hamilton, Koehler, and Moitra [HKM17] proposed a greedy algorithm to learn pairwise (as well as higher-order) MRFs with general alphabet. Their algorithm generalizes Bresler's approach for learning Ising models [Bre15]. The sample complexity in [HKM17] grows logarithmically in n, but *doubly* exponentially in the width λ (only single exponential is necessary for learning Ising models [SW12]). Klivans and Meka [KM17] provided a different algorithmic and theoretical approach by setting this up as an online learning problem and leveraging results from the Hedge algorithm therein. Their algorithm Sparsitron achieves single-exponential dependence on the width λ.

Our Contributions

- Our main result: We show that the $\ell_{2,1}$-constrained logistic regression can recover the underlying graph from i.i.d. samples of a discrete pairwise graphical model. For the special case of Ising models, this reduces to an ℓ_1-constrained logistic regression. We make no incoherence assumption on the graph structure. As shown in Table 1, our sample complexity scales as $\tilde{O}(k^4)$, which improves the previous best result with $\tilde{O}(k^5)$ dependency. The analysis applies a sharp generalization error bound for logistic regression when the weight vector has an $\ell_{2,1}$ constraint (or ℓ_1 constraint) and the sample vector has an $\ell_{2,\infty}$ constraint (or ℓ_∞ constraint). Our key insight is that a generalization bound can be used to control the squared distance between the predicted and true logistic functions, which then implies an ℓ_∞ norm bound between the weight vectors.

- We show that the proposed algorithms can run in $\tilde{O}(n^2)$ time without affecting the statistical guarantees (see Section 2.3). Note that $\tilde{O}(n^2)$ is an efficient runtime for graph recovery over n nodes. Previous algorithms in [HKM17, KM17] require $\tilde{O}(n^2)$ runtime for learning pairwise graphical models.

- We construct examples that violate the incoherence condition proposed in [RWL10] (see Figure 1). We then run ℓ_1-constrained logistic regression and show that it can recover the graph structure as long as given enough samples. This verifies our analysis and also shows that our conditions for graph recovery are weaker than those in [RWL10].

- We empirically compare the proposed algorithm with the Sparsitron algorithm in [KM17] over different alphabet sizes, and show that our algorithm needs fewer samples for graph recovery (see Figure 3).

Notation. We use $[n]$ to denote the set $\{1, 2, \cdots, n\}$. For a vector $x \in \mathbb{R}^n$, we use x_i or $x(i)$ to denote its i-th coordinate. The ℓ_p norm of a vector is defined as $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$. We use $x_{-i} \in \mathbb{R}^{n-1}$ to denote the vector after deleting the i-th coordinate. For matrix $A \in \mathbb{R}^{n \times k}$, we use $A(i, j)$ to denote its (i, j)-th entry. We use $A(i, :) \in \mathbb{R}^k$ and $A(:, j) \in \mathbb{R}^n$ to the denote the

i-th row vector and the j-th column vector. The $\ell_{p,q}$ norm of a matrix $A \in \mathbb{R}^{n\times k}$ is defined as $\|A\|_{p,q} = \|[\|A(1,:)\|_p, ..., \|A(n,:)\|_p]\|_q$. We define $\|A\|_\infty = \max_{ij} |A(i,j)|$ throughout the paper (note that this definition is different from the induced matrix norm). We use $\sigma(z) = 1/(1+e^{-z})$ to represent the sigmoid function. We use $\langle \cdot, \cdot \rangle$ to represent the dot product between two vectors $\langle x, y\rangle = \sum_i x_i y_i$ or two matrices $\langle A, B\rangle = \sum_{ij} A(i,j)B(i,j)$.

2 Main results

We start with the special case of binary variables (i.e., Ising models), and then move to the general setting with non-binary variables.

2.1 Learning Ising models

We first give a formal definition of an Ising model distribution.

Definition 1. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric weight matrix with $A_{ii} = 0$ for $i \in [n]$. Let $\theta \in \mathbb{R}^n$ be a mean-field vector. The n-variable Ising model is a distribution $\mathcal{D}(A,\theta)$ on $\{-1,1\}^n$ that satisfies*

$$\mathbb{P}_{Z\sim\mathcal{D}(A,\theta)}[Z = z] \propto \exp(\sum_{1\le i<j\le n} A_{ij} z_i z_j + \sum_{i\in[n]} \theta_i z_i). \tag{1}$$

The dependency graph of $\mathcal{D}(A,\theta)$ is an undirected graph $G = (V,E)$, with vertices $V = [n]$ and edges $E = \{(i,j) : A_{ij} \neq 0\}$. The width of $\mathcal{D}(A,\theta)$ is defined as

$$\lambda(A,\theta) = \max_{i\in[n]}(\sum_{j\in[n]} |A_{ij}| + |\theta_i|). \tag{2}$$

Let $\eta(A,\theta)$ be the minimum edge weight in absolute value, i.e., $\eta(A,\theta) = \min_{i,j\in[n]:A_{ij}\neq 0} |A_{ij}|$.

One important property of an Ising model distribution is that the conditional distribution of any variable given the rest variables follows a logistic function. Let $\sigma(z) = 1/(1+e^{-z})$ be the sigmoid function.

Fact 1. *Let $Z \sim \mathcal{D}(A,\theta)$ and $Z \in \{-1,1\}^n$. For any $i \in [n]$, the conditional probability of the i-th variable $Z_i \in \{-1,1\}$ given the states of all other variables $Z_{-i} \in \{-1,1\}^{n-1}$ is*

$$\mathbb{P}[Z_i = 1|Z_{-i} = x] = \frac{\exp(\sum_{j\neq i} A_{ij}x_j + \theta_i)}{\exp(\sum_{j\neq i} A_{ij}x_j + \theta_i) + \exp(-\sum_{j\neq i} A_{ij}x_j - \theta_i)} = \sigma(\langle w, x'\rangle), \tag{3}$$

where $x' = [x, 1] \in \{-1,1\}^n$, and $w = 2[A_{i1}, \cdots, A_{i(i-1)}, A_{i(i+1)}, \cdots, A_{in}, \theta_i] \in \mathbb{R}^n$. Moreover, w satisfies $\|w\|_1 \le 2\lambda(A,\theta)$, where $\lambda(A,\theta)$ is the model width defined in Definition 1.

Given N i.i.d. samples $\{z^1, \cdots, z^N\}$, $z^i \in \{-1,1\}^n$ from an Ising model $\mathcal{D}(A,\theta)$, one natural approach to graph recovery is to form a logistic regression problem for each variable separately, and then recover the associated edge weights. For simplicity, let us focus on the n-th variable (the algorithm directly applies to the rest variables). We first transform the samples into $\{(x^i, y^i)\}_{i=1}^N$, where $x^i = [z_1^i, \cdots, z_{n-1}^i, 1] \in \{-1,1\}^n$ and $y^i = z_n^i \in \{-1,1\}$. By Fact 1, we know that $\mathbb{P}[y^i = 1|x^i = x] = \sigma(\langle w^*, x\rangle)$ where $w^* = 2[A_{n1}, \cdots, A_{n(n-1)}, \theta_n] \in \mathbb{R}^n$

Algorithm 1 Learning Ising model via ℓ_1-constrained logistic regression

Input: N i.i.d. samples $\{z^1, \cdots, z^N\}$, $z^m \in \{-1,1\}^n$, for $m \in [N]$; an upper bound on $\lambda(A,\theta) \le \lambda$; a lower bound on $\eta(A,\theta) \ge \eta > 0$.
Output: $\hat{A} \in \mathbb{R}^{n\times n}$, and an undirected graph $\hat{G}$ on n nodes.

```
for each node $i \in [n]$ do
    $\forall m \in [N]$, $x^m \leftarrow [z^m_{-i}, 1]$, $y^m \leftarrow z^m_i$    $\triangleright$ Form samples $(x^m, y^m) \in \{-1,1\}^n \times \{-1,1\}$.
    $\hat{w} \leftarrow \arg\min_{w\in\mathbb{R}^n} \frac{1}{N}\sum_{m=1}^N \ln(1+e^{-y^m\langle w,x^m\rangle})$   s.t. $\|w\|_1 \le 2\lambda$.
    $\forall j \in [n]$, $\hat{A}_{ij} \leftarrow \hat{w}_{\tilde{j}}/2$, where $\tilde{j} = j$ if $j < i$ and $\tilde{j} = j-1$ if $j > i$.
Form an undirected graph $\hat{G}$ on $n$ nodes with edges $\{(i,j) : |\hat{A}_{ij}| \ge \eta/2, i < j\}$.
```

satisfies $\|w^*\|_1 \le 2\lambda(A,\theta)$. Suppose that $\lambda(A,\theta) \le \lambda$, we are then interested in recovering w^* by the following ℓ_1-constrained logistic regression problem

$$\hat{w} \in \min_{w\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^N \ell(y^i \langle w, x^i\rangle) \qquad \text{s.t. } \|w\|_1 \le 2\lambda, \tag{4}$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is the loss function

$$\ell(y^i \langle w, x^i\rangle) = \ln(1+e^{-y^i\langle w,x^i\rangle}) = \begin{cases} -\ln\sigma(\langle w, x^i\rangle), & \text{if } y^i = 1 \\ -\ln(1-\sigma(\langle w, x^i\rangle)), & \text{if } y^i = -1 \end{cases} \tag{5}$$

Eq. (5) is essentially the negative log-likelihood of observing y^i given x^i at the current w.

Let $\hat{w}$ be a minimizer of (4). It is worth noting that in the high-dimensional regime ($N < n$), $\hat{w}$ may not be unique[1]. In this case, we will show that *any* one of them would work. The edge weight is estimated as $\hat{A}_{nj} = \hat{w}_j/2$.

The pseudocode of the above algorithm is given in Algorithm 1. Solving the ℓ_1-constrained logistic regression problem will give us an estimator of the true edge weight. We then form the graph by keeping the edge that has estimated weight larger than $\eta/2$ (in absolute value).

Theorem 1. *Let $\mathcal{D}(A,\theta)$ be an unknown n-variable Ising model distribution with dependency graph G. Suppose that the $\mathcal{D}(A,\theta)$ has width $\lambda(A,\theta) \le \lambda$. Given $\rho \in (0,1)$ and $\epsilon > 0$, if the number of i.i.d. samples satisfies $N = O(\lambda^2 \exp(12\lambda)\ln(n/\rho)/\epsilon^4)$, then with probability at least $1-\rho$, Algorithm 1 produces $\hat{A}$ that satisfies*

$$\max_{i,j\in[n]} |A_{ij} - \hat{A}_{ij}| \le \epsilon. \tag{6}$$

Corollary 1. *In the setup of Theorem 1, suppose that the Ising model distribution $\mathcal{D}(A,\theta)$ has minimum edge weight $\eta(A,\theta) \ge \eta > 0$. If we set $\epsilon < \eta/2$ in (6), which corresponds to sample complexity $N = O(\lambda^2\exp(12\lambda)\ln(n/\rho)/\eta^4)$, then with probability at least $1-\rho$, Algorithm 1 recovers the dependency graph, i.e., $\hat{G} = G$.*

[1] But they all have the same loss function value since (4) is a convex program.

2.2 Learning pairwise graphical models over general alphabet

We first give a formal definition of the general pairwise graphical model.

Definition 2. *Let k be the alphabet size. Let $\mathcal{W} = \{W_{ij} \in \mathbb{R}^{k\times k} : i \neq j \in [n]\}$ be a set of weight matrices satisfying $W_{ij} = W_{ji}^T$. Without loss of generality, we assume that for any $i \neq j$, each row vector as well as the column vector of W_{ij} has zero mean. Let $\Theta = \{\theta_i \in \mathbb{R}^k : i \in [n]\}$ be a set of external field vectors. Then the n-variable pairwise graphical model $\mathcal{D}(\mathcal{W}, \Theta)$ is a distribution over $[k]^n$ where*

$$\mathop{\mathbb{P}}_{Z\sim\mathcal{D}(\mathcal{W},\Theta)}[Z = z] \propto \exp(\sum_{1\le i<j\le n} W_{ij}(z_i, z_j) + \sum_{i\in[n]} \theta_i(z_i)). \tag{7}$$

The dependency graph of $\mathcal{D}(\mathcal{W}, \Theta)$ is an undirected graph $G = (V, E)$, with vertices $V = [n]$ and edges $E = \{(i, j) : W_{ij} \neq 0\}$. The width of $\mathcal{D}(\mathcal{W}, \Theta)$ is defined as

$$\lambda(\mathcal{W}, \Theta) = \max_{i,a}(\sum_{j\neq i} \max_{b\in[k]} |W_{ij}(a, b)| + |\theta_i(a)|). \tag{8}$$

Define $\eta(\mathcal{W}, \Theta)$ as

$$\eta(\mathcal{W}, \Theta) = \min_{(i,j)\in E} \max_{a,b} |W_{ij}(a, b)|. \tag{9}$$

Remark. The assumption that W_{ij} has centered rows and columns (i.e., $\sum_b W_{ij}(a, b) = 0$ and $\sum_a W_{ij}(a, b) = 0$ for any $a, b \in [k]$) is without loss of generality (see Fact 8.2 in [KM17]). If the a-th row of W_{ij} is not centered, i.e., $\sum_b W_{ij}(a, b) \neq 0$, we can define $W'_{ij}(a, b) = W_{ij}(a, b) - \sum_b W_{ij}(a, b)/k$ and $\theta'_i(a) = \theta_i(a) + \sum_b W_{ij}(a, b)/k$, and notice that $\mathcal{D}(\mathcal{W}, \Theta) = \mathcal{D}(\mathcal{W}', \Theta')$. Because the sets of matrices with centered rows and columns (i.e., $\{W_{ij}, 1 \le i < j \le n : \sum_b W_{ij}(a, b) = 0, \forall a \in [k]\}$ and $\{W_{ij}, 1 \le i < j \le n : \sum_a W_{ij}(a, b) = 0, \forall b \in [k]\}$) are two linear subspaces, alternatively projecting W_{ij} onto the two sets will converge to the intersection of the two subspaces [VN49]. According to the previous discussion, the condition of centered rows and columns is necessary for recovering the underlying weight matrices, since otherwise different parameters can give the same distribution.

For a pairwise graphical model distribution $\mathcal{D}(\mathcal{W}, \Theta)$, the conditional distribution of any variable (when restricted to a pair of values) given all the other variables follows a logistic function, as shown in Fact 2. This is analogous to Fact 1 for the Ising model distribution.

Fact 2. *Let $Z \sim \mathcal{D}(\mathcal{W}, \Theta)$ and $Z \in [k]^n$. For any $i \in [n]$, and any $\alpha \neq \beta \in [k]$, we have*

$$\mathbb{P}[Z_i = \alpha | Z_i \in \{\alpha, \beta\}, Z_{-i} = x] = \sigma(\sum_{j\neq i}(W_{ij}(\alpha, x_j) - W_{ij}(\beta, x_j)) + \theta_i(\alpha) - \theta_i(\beta)). \tag{10}$$

Given N i.i.d. samples $\{z^1, \cdots, z^N\}$, where $z^i \in [k]^n \sim \mathcal{D}(\mathcal{W}, \Theta)$, the goal is to estimate matrices W_{ij} for all $i \neq j \in [n]$. For simplicity, let us focus on the n-th variable (the algorithm directly extends to other variables). Now the goal is to estimate matrices W_{nj} for all $j \in [n-1]$.

To use Fact 2, fix a pair of values $\alpha \neq \beta \in [k]$, let S be the set of samples satisfying $z_n \in \{\alpha, \beta\}$. We next transform the samples in S to $\{(x^i, y^i)\}_{i=1}^{|S|}$ as follows: $x^i = \text{OneHotEncode}([z^i_{-n}, 1]) \in \{0, 1\}^{n\times k}$, $y^i = 1$ if $z^i_n = \alpha$, and $y^i = -1$ if $z^i_n = \beta$. Here

OneHotEncode$(\cdot) : [k]^n \to \{0,1\}^{n\times k}$ is a function that maps a value $i \in [k]$ to the standard basis vector $e_i \in \{0,1\}^k$, i.e., e_i has a single 1 at the i-th entry.

For samples $\{(x^i, y^i)\}_{i=1}^{|S|}$ in set S, Fact 2 implies that $\mathbb{P}[y = 1|x] = \sigma(\langle w^*, x\rangle)$, where $w^* \in \mathbb{R}^{n\times k}$ satisfies

$$w^*(j,:) = W_{nj}(\alpha,:) - W_{nj}(\beta,:), \forall j \in [n-1]; \quad w^*(n,:) = [\theta_i(\alpha) - \theta_i(\beta), 0, ..., 0]. \tag{11}$$

Suppose that the width of $\mathcal{D}(\mathcal{W},\Theta)$ satisfies $\lambda(\mathcal{W},\Theta) \le \lambda$, then w^* defined in (11) satisfies $\|w^*\|_{2,1} \le 2\lambda\sqrt{k}$, where $\|w^*\|_{2,1} := \sum_j \|w^*(j,:)\|_2$. We can now form an $\ell_{2,1}$-constrained logistic regression over the samples in S:

$$w^{\alpha,\beta} \in \arg\min_{w\in\mathbb{R}^{n\times k}} \frac{1}{|S|}\sum_{i=1}^{|S|} \ln(1 + e^{-y^i\langle w, x^i\rangle}) \qquad \text{s.t. } \|w\|_{2,1} \le 2\lambda\sqrt{k}, \tag{12}$$

Let $w^{\alpha,\beta}$ be a minimizer of (12). Without loss of generality, we can assume that the first $n-1$ rows of $w^{\alpha,\beta}$ are centered, i.e., $\sum_a w^{\alpha,\beta}(j,a) = 0$ for $j \in [n-1]$. Otherwise, we can always define a new matrix $U^{\alpha,\beta} \in \mathbb{R}^{n\times k}$ by centering the first $n-1$ rows of $w_{\alpha,\beta}$:

$$U^{\alpha,\beta}(j,:) = w^{\alpha,\beta}(j,:) - \frac{1}{k}\sum_{a\in[k]} w^{\alpha,\beta}(j,a),\ \forall j \in [n-1]; \tag{13}$$

$$U^{\alpha,\beta}(n,:) = w^{\alpha,\beta}(n,:) + \frac{1}{k}\sum_{j\in[n-1], a\in[k]} w^{\alpha,\beta}(j,a).$$

Since each row of the x matrix in (12) is a standard basis vector, $\langle U^{\alpha,\beta}, x\rangle = \langle w^{\alpha,\beta}, x\rangle$, which implies that $U^{\alpha,\beta}$ is also a minimizer of (12).

The key step in the proof is to show that given enough samples, the obtained $U^{\alpha,\beta} \in \mathbb{R}^{n\times k}$ matrix is close to $W_{nj}(\alpha,:) - W_{nj}(\beta,:)$ in absolute distance:

$$|W_{nj}(\alpha,:) - W_{nj}(\beta,:) - U^{\alpha,\beta}(j,:)| \le \epsilon, \quad \forall j \in [n-1],\ \forall \alpha,\beta \in [k]. \tag{14}$$

Recall that our goal is to estimate the original matrices W_{nj} for all $j \in [n-1]$. Summing (14) over $\beta \in [k]$ (suppose $U^{\alpha,\alpha} = 0$) and using the fact that $\sum_\beta W_{nj}(\beta,:) = 0$ gives

$$|W_{nj}(\alpha,:) - \frac{1}{k}\sum_{\beta\in[k]} U^{\alpha,\beta}(j,:)| \le \epsilon, \quad \forall j \in [n-1],\ \forall \alpha \in [k]. \tag{15}$$

In other words, $\hat{W}_{nj}(\alpha,:) = \frac{1}{k}\sum_{\beta\in[k]} U^{\alpha,\beta}(j,:)$ is a good estimate of $W_{nj}(\alpha,:)$.

Suppose that $\eta(\mathcal{W},\Theta) \ge \eta$, once we obtain the estimates $\hat{W}_{ij}$, the last step is to form a graph by keeping the edge (i,j) that satisfies $\max_{a,b} |\hat{W}_{ij}(a,b)| \ge \eta/2$. The pseudocode of the above algorithm is given in Algorithm 2.

Theorem 2. *Let $\mathcal{D}(\mathcal{W},\Theta)$ be an n-variable pairwise graphical model distribution with width $\lambda(\mathcal{W},\Theta) \le \lambda$. Given $\rho \in (0,1)$ and $\epsilon > 0$, if the number of i.i.d. samples satisfies $N = O(\lambda^2 k^4 \exp(14\lambda)\ln(nk/\rho)/\epsilon^4)$, then with probability at least $1-\rho$, Algorithm 2 produces $\hat{W}_{ij} \in \mathbb{R}^{k\times k}$ that satisfies*

$$|W_{ij}(a,b) - \hat{W}_{ij}(a,b)| \le \epsilon, \quad \forall i \ne j \in [n],\ \forall a,b \in [k]. \tag{16}$$

Algorithm 2 Learning pairwise graphical models via $\ell_{2,1}$-constrained logistic regression

Input: alphabet size k; N i.i.d. samples $\{z^1, \cdots, z^N\}$, where $z^m \in [k]^n$ for $m \in [N]$; an upper bound on $\lambda(\mathcal{W}, \Theta) \leq \lambda$; a lower bound on $\eta(\mathcal{W}, \Theta) \geq \eta > 0$.
Output: $\hat{W}_{ij} \in \mathbb{R}^{k \times k}$ for all $i \neq j \in [n]$; an undirected graph $\hat{G}$ on n nodes.

```
for each node $i \in [n]$ do
    for each pair $\alpha \neq \beta \in [k]$ do
        $S \leftarrow \{z^m, m \in [N] : z_i^m \in \{\alpha, \beta\}\}$
        $\forall z^t \in S$, $x^t \leftarrow \mathrm{OneHotEncode}([z_{-i}^t, 1])$, $y^t \leftarrow 1$ if $z_i^t = \alpha$; $y^t \leftarrow -1$ if $z_i^t = \beta$.
        $w^{\alpha,\beta} \leftarrow \arg\min_{w \in \mathbb{R}^{n \times k}} \frac{1}{|S|} \sum_{t=1}^{|S|} \ln(1 + e^{-y^t \langle w, x^t \rangle})$  s.t. $\|w\|_{2,1} \leq 2\lambda\sqrt{k}$
        Define $U^{\alpha,\beta} \in \mathbb{R}^{n \times k}$ by centering the first $n-1$ rows of $w^{\alpha,\beta}$ (see (13)).
    for $j \in [n] \backslash i$ and $\alpha \in [k]$ do
        $\hat{W}_{ij}(\alpha, :) = \frac{1}{k} \sum_{\beta \in [k]} U^{\alpha,\beta}(\tilde{j}, :)$, where $\tilde{j} = j$ if $j < i$ and $\tilde{j} = j - 1$ if $j > i$.
Form graph $\hat{G}$ on $n$ nodes with edges $\{(i,j) : \max_{a,b} |\hat{W}_{ij}(a,b)| \geq \eta/2, i < j\}$.
```

Corollary 2. *In the setup of Theorem 2, suppose that the pairwise graphical model distribution $\mathcal{D}(\mathcal{W}, \Theta)$ satisfies $\eta(\mathcal{W}, \Theta) \geq \eta > 0$. If we set $\epsilon < \eta/2$ in Theorem 2, which corresponds to sample complexity $N = O(\lambda^2 k^4 \exp(14\lambda) \ln(nk/\rho)/\eta^4)$, then with probability at least $1 - \rho$, Algorithm 2 recovers the dependency graph, i.e., $\hat{G} = G$.*

Remark. The $w^* \in \mathbb{R}^{n \times k}$ matrix defined in (11) satisfies $\|w^*\|_{\infty,1} \leq 2\lambda(\mathcal{W}, \Theta)$. This implies that $\|w^*\|_{2,1} \leq 2\lambda(\mathcal{W}, \Theta)\sqrt{k}$ and $\|w^*\|_1 \leq 2\lambda(\mathcal{W}, \Theta)k$. Instead of solving the $\ell_{2,1}$-constrained logistic regression defined in (12), we could solve an ℓ_1-constrained logistic regression with $\|w\|_1 \leq 2\lambda(\mathcal{W}, \Theta)k$. However, this will lead to a sample complexity that scales as $\tilde{O}(k^5)$, which is worse than the $\tilde{O}(k^4)$ sample complexity achieved by the $\ell_{2,1}$-constrained logistic regression. The reason why we use the $\ell_{2,1}$ constraint instead of the tighter $\ell_{\infty,1}$ constraint in the algorithm is because our proof relies on a sharp generalization bound for $\ell_{2,1}$-constrained logistic regression (see Lemma 10 in the appendix). It is unclear whether a similar generalization bound exists for the $\ell_{\infty,1}$ constraint.

2.3 Learning pairwise graphical models in $\tilde{O}(n^2)$ time

Our results so far assume that the ℓ_1-constrained logistic regression (in Algorithm 1) and the $\ell_{2,1}$-constrained logistic regression (in Algorithm 2) can be solved exactly. This would require $\tilde{O}(n^4)$ complexity if an interior-point based method is used [KKB07]. The goal of this section is to reduce the runtime to $\tilde{O}(n^2)$ via first-order optimization method. Note that $\tilde{O}(n^2)$ is an efficient time complexity for graph recovery over n nodes. Previous structural learning algorithms of Ising models require either $\tilde{O}(n^2)$ complexity (e.g., [Bre15, KM17]) or a worse complexity (e.g., [RWL10, VMLC16] require $\tilde{O}(n^4)$ runtime[2]). We would like to remark that our goal of this section is not to give the fastest first-order optimization algorithm (see the discussion after Theorem 4). Instead, our contribution here is to provably show that it is

[2] If we change the ℓ_1-regularized interaction screening estimator proposed in [VMLC16] to an ℓ_1-constrained version, then it may be possible to apply the proposed mirror descent algorithm to optimize it. The main problem is that now the original proof of [VMLC16] does not work, and one needs a new proof for the ℓ_1-constrained interaction screening estimator.

possible to run Algorithm 1 and Algorithm 2 in $\tilde{O}(n^2)$ time without affecting the original statistical guarantees.

To better exploit the problem structure[3], we use the mirror descent algorithm[4] with a properly chosen distance generating function (aka the mirror map). Following the standard mirror descent setup, we use negative entropy as the mirror map for ℓ_1-constrained logistic regression and a scaled group norm for $\ell_{2,1}$-constrained logistic regression (see Section 5.3.3.2 and Section 5.3.3.3 in [BTN13] for more details). The pseudocode is given in Appendix H. The main advantage of mirror descent algorithm is that its convergence rate scales logarithmically in the dimension. Specifically, let $\bar{w}$ be the output after $O(\ln(n)/\gamma^2)$ mirror descent iterations, then $\bar{w}$ satisfies

$$\hat{\mathcal{L}}(\bar{w}) - \hat{\mathcal{L}}(\hat{w}) \leq \gamma, \tag{17}$$

where $\hat{\mathcal{L}}(w) = \sum_{i=1}^{N} \ln(1 + e^{-y^i \langle w, x^i \rangle})/N$ is the empirical logistic loss, and $\hat{w}$ is the actual minimizer of $\hat{\mathcal{L}}(w)$. Since each mirror descent update requires $O(nN)$ time, where N is the number of samples and scales as $O(\ln(n))$, and we have to solve $O(n)$ regression problems for n variables, the total runtime scales as $\tilde{O}(n^2)$.

There is still one problem left, that is, we have to show that $\|\bar{w} - w^*\|_\infty \leq \epsilon$ (where w^* is the minimizer of the true loss $\mathcal{L}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \ln(1 + e^{-y\langle w, x\rangle}))$ in order to conclude that (6) and (16) hold when using mirror descent algorithms. Since $\hat{\mathcal{L}}(w)$ is not strongly convex, (17) alone does not necessarily imply that $\|\bar{w} - \hat{w}\|_\infty$ is small. Fortunately, we note that in the proof of Theorem 1 and Theorem 2, the definition of $\hat{w}$ (as a minimizer of $\hat{\mathcal{L}}(w)$) is only used to show that $\hat{\mathcal{L}}(\hat{w}) \leq \hat{\mathcal{L}}(w^*)$. It is then possible to replace it with (17) in the original proof, and prove that Theorem 1 and Theorem 2 still hold as long as γ is small enough.

Our key results in this section are Theorem 3 and Theorem 4, which show that Algorithm 1 and Algorithm 2 can be used to recover the dependency graph in $\tilde{O}(n^2)$ time.

Theorem 3. *In the setup of Theorem 1, suppose that the ℓ_1-constrained logistic regression in Algorithm 1 is optimized using the mirror descent algorithm given in Appendix H. Given $\rho \in (0,1)$ and $\epsilon > 0$, if the number of mirror descent iterations satisfies $T = O(\lambda^2 \exp(12\lambda)\ln(n)/\epsilon^4)$, and the number of i.i.d. samples satisfies $N = O(\lambda^2 \exp(12\lambda)\ln(n/\rho)/\epsilon^4)$, then (6) still holds with probability at least $1-\rho$. The total time complexity of Algorithm 1 is $O(TNn^2)$.*

Theorem 4. *In the setup of Theorem 2, suppose that the $\ell_{2,1}$-constrained logistic regression in Algorithm 2 is optimized using the mirror descent algorithm given in Appendix H. Given $\rho \in (0,1)$ and $\epsilon > 0$, if the number of mirror descent iterations satisfies $T = O(\lambda^2 k^3 \exp(12\lambda)\ln(n)/\epsilon^4)$, and the number of i.i.d. samples satisfies $N = O(\lambda^2 k^4 \exp(14\lambda)\ln(nk/\rho)/\epsilon^4)$, then (16) still holds with probability at least $1-\rho$. The total time complexity of Algorithm 2 is $O(TNn^2k^2)$.*

Remark. The proposed algorithms can be easily parallelized since the logistic regression is defined separately for each variable. Besides, it is possible to improve the time complexity

[3] Specifically, for the ℓ_1-constrained logisitic regression defined in (4), since the input sample satisifies $\|x\|_\infty = 1$, the loss function is $O(1)$-Lipschitz w.r.t. $\|\cdot\|_1$. Similarly, for the $\ell_{2,1}$-constrained logisitic regression defined in (12), the loss function is $O(1)$-Lipschitz w.r.t. $\|\cdot\|_{2,1}$ because the input sample satisifies $\|x\|_{2,\infty} = 1$.

[4] Other approaches include the standard projected gradient descent and the coordinate descent. Their convergence rates depend on either the smoothness or the Lipschitz constant (w.r.t. $\|\cdot\|_2$) of the objective function [Bub15]. This would lead to a total runtime of $\tilde{O}(n^3)$ for our problem setting. Another option would be the composite gradient descent method, the analysis of which relies on the restriced strong convexity of the objective function [ANW10]. For other variants of mirror descent algorithms, see the remark after Theorem 4.

given in Theorem 1 and Theorem 2 (especially the dependence on ϵ and λ), by using stochastic or accelerated versions of mirror descent algorithms (instead of the batch version given in Appendix H). For example, if online mirror descent algorithms are used, then the runtime would be $O(Nn^2)$ and $O(Nn^2k^2)$ simply because each mirror descent update uses a single sample instead of all samples (and the number of updates equals the number of samples). In fact, the Sparsitron algorithm proposed by Klivans and Meka [KM17] can be seen as an online mirror descent algorithm for optimizing ℓ_1-constrained logistic regression (see Algorithm 3 given in Appendix H). As pointed out at the beginning of this section, our goal here is not to give the most efficient optimization algorithm. The focus of this section is to show that it is possible to run Algorithm 1 and Algorithm 2 in $\tilde{O}(n^2)$ time and achieve the same statistical guarantee.

3 Analysis

3.1 Proof outline

We give a proof outline for learning Ising models (the general setting with non-binary alphabet follows a similar outline). Let D be a distribution over $\{-1,1\}^n \times \{-1,1\}$, where $(x,y) \sim D$ satisfies $\mathbb{P}[y = 1|x] = \sigma(\langle w^*, x\rangle)$. Let $\mathcal{L}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \ln(1 + e^{-y\langle w,x\rangle})$ and $\hat{\mathcal{L}}(w) = \sum_{i=1}^N \ln(1 + e^{-y^i\langle w,x^i\rangle})/N$ be the expected and empirical logistic loss. Suppose $\|w^*\|_1 \le 2\lambda$. Let $\hat{w} \in \arg\min_w \hat{\mathcal{L}}(w)$ s.t. $\|w\|_1 \le 2\lambda$. Our proof can be summarized in three steps:

1. If the number of samples satisfies $N = O(\lambda^2 \ln(n/\rho)/\gamma^2)$, then $\mathcal{L}(\hat{w}) - \mathcal{L}(w^*) \le O(\gamma)$. This is obtained using a sharp generalization bound when $\|w\|_1$ and $\|x\|_\infty$ are bounded (see Lemma 7 in Appendix B).

2. For any w, we show that $\mathcal{L}(w) - \mathcal{L}(w^*) \ge \mathbb{E}_x[\sigma(\langle w, x\rangle) - \sigma(\langle w^*, x\rangle)]^2$ (see Lemma 9 and Lemma 8 in Appendix B). Hence, Step 1 implies that $\mathbb{E}_x[\sigma(\langle \hat{w}, x\rangle) - \sigma(\langle w^*, x\rangle)]^2 \le O(\gamma)$ (see Lemma 1 in the next subsection).

3. We now use a result from [KM17] (see Lemma 5 in the next subsection) to show that if the samples are from an Ising model and $\gamma = O(\epsilon^2 \exp(-6\lambda))$, then $\mathbb{E}_x[\sigma(\langle \hat{w}, x\rangle) - \sigma(\langle w^*, x\rangle)]^2 \le O(\gamma)$ implies that $\|\hat{w} - w^*\|_\infty \le \epsilon$.

For the general setting with non-binary alphabet, the proof follows a similar outline. The main difference is that we need to use a sharp generalization bound when $\|w\|_{2,1}$ and $\|x\|_{2,\infty}$ are bounded (see Lemma 10 in Appendix B). This would give us Lemma 2 (instead of Lemma 1 for the Ising model setting). The last step is to use Lemma 6 to bound the infinity norm between the two weight matrices.

3.2 Supporting lemmas

Lemma 1 and Lemma 2 are the key results in our proof. They essentially say that given enough samples, solving the corresponding constrained logistic regression problem will provide a prediction $\sigma(\langle \hat{w}, x\rangle)$ close to the true $\sigma(\langle w^*, x\rangle)$ in terms of their expected squared distance.

Lemma 1. *Let $\mathcal{D}$ be a distribution on $\{-1,1\}^n \times \{-1,1\}$ where for $(X,Y) \sim \mathcal{D}$, $\mathbb{P}[Y=1|X=x] = \sigma(\langle w^*, x\rangle)$. We assume that $\|w^*\|_1 \le 2\lambda$ for a known $\lambda \ge 0$. Given N i.i.d. samples $\{(x^i, y^i)\}_{i=1}^N$, let $\hat{w}$ be any minimizer of the following ℓ_1-constrained logistic regression problem:*

$$\hat{w} \in \arg\min_{w \in \mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N} \ln(1+e^{-y^i\langle w, x^i\rangle}) \quad \text{s.t. } \|w\|_1 \le 2\lambda. \tag{18}$$

Given $\rho \in (0,1)$ and $\epsilon > 0$, suppose that $N = O(\lambda^2(\ln(n/\rho))/\epsilon^2)$, then with probability at least $1-\rho$ over the samples, we have that $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(\langle w^*, x\rangle) - \sigma(\langle \hat{w}, x\rangle))^2] \le \epsilon$.

Lemma 2. *Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{-1,1\}$, where $\mathcal{X} = \{x \in \{0,1\}^{n\times k} : \|x\|_{2,\infty} \le 1\}$. Furthermore, $(X,Y) \sim \mathcal{D}$ satisfies $\mathbb{P}[Y=1|X=x] = \sigma(\langle w^*, x\rangle)$, where $w^* \in \mathbb{R}^{n\times k}$. We assume that $\|w^*\|_{2,1} \le 2\lambda\sqrt{k}$ for a known $\lambda \ge 0$. Given N i.i.d. samples $\{(x^i, y^i)\}_{i=1}^N$ from $\mathcal{D}$, let $\hat{w}$ be any minimizer of the following $\ell_{2,1}$-constrained logistic regression problem:*

$$\hat{w} \in \arg\min_{w \in \mathbb{R}^{n\times k}} \frac{1}{N}\sum_{i=1}^{N} \ln(1+e^{-y^i\langle w, x^i\rangle}) \quad \text{s.t. } \|w\|_{2,1} \le 2\lambda\sqrt{k}. \tag{19}$$

Given $\rho \in (0,1)$ and $\epsilon > 0$, suppose that $N = O(\lambda^2 k(\ln(n/\rho))/\epsilon^2)$, then with probability at least $1-\rho$ over the samples, we have that $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(\langle w^*, x\rangle) - \sigma(\langle \hat{w}, x\rangle))^2] \le \epsilon$.

The proofs of Lemma 1 and Lemma 2 are given in Appendix B. Note that in the setup of both lemmas, we form a pair of dual norms for x and w, e.g., $\|x\|_{2,\infty}$ and $\|w\|_{2,1}$ in Lemma 2, and $\|x\|_\infty$ and $\|w\|_1$ in Lemma 1. This duality allows us to use a sharp generalization bound with a sample complexity that scales logarithmic in the dimension.

Intuitively, if a variable in a graphical model distribution concentrates on a subset of the alphabet (e.g., it always takes the same value in an Ising model distribution), then it is difficult to infer the exact relation between this variable and other variables. One key property of the graphical model distribution is that this bad event cannot happen. The (conditional) probability that a variable takes any value in the alphabet is lowered bounded by a nonzero quantity (see Definition 3 and Lemma 4).

Definition 3. *Let S be the alphabet set, e.g., $S = \{-1,1\}$ for Ising model and $S = [k]$ for an alphabet of size k. A distribution $\mathcal{D}$ on S^n is δ-unbiased if for $X \sim \mathcal{D}$, any $i \in [n]$, and any assignment $x \in S^{n-1}$ to X_{-i}, $\min_{\alpha\in S}(\mathbb{P}[X_i = \alpha | X_{-i} = x]) \ge \delta$.*

This notion of δ-unbiased distribution is proposed by Klivans and Meka [KM17]. For a δ-unbiased distribution, any of its marginal distribution is also δ-unbiased, as indicated by the following lemma.

Lemma 3. *Let $\mathcal{D}$ be a δ-unbiased distribution on S^n, where S is the alphabet set. For $X \sim \mathcal{D}$, any $i \in [n]$, the distribution of X_{-i} is also δ-unbiased.*

Lemma 4 describes the δ-unbiased property of MRFs. This property has been used in the previous papers (e.g., [KM17, Bre15]). For completeness, we also give a proof of Lemma 3 and Lemma 4 in the appendix.

Lemma 4. *Let $\mathcal{D}(\mathcal{W}, \Theta)$ be a pairwise graphical model distribution with alphabet size k and width $\lambda(\mathcal{W}, \Theta)$. Then $\mathcal{D}(\mathcal{W}, \Theta)$ is δ-unbiased with $\delta = e^{-2\lambda(\mathcal{W},\Theta)}/k$. Specifically, an Ising model distribution $\mathcal{D}(A, \theta)$ is $e^{-2\lambda(A,\theta)}/2$-unbiased.*

Recall that Lemma 1 and Lemma 2 give a sample complexity bound for achieving a small ℓ_2 error between $\sigma(\langle \hat{w}, x\rangle)$ and $\sigma(\langle w^*, x\rangle)$. We still need to show that $\hat{w}$ is close to w^*. The following two lemmas provide a connection between the ℓ_2 error and $\|\hat{w} - w^*\|_\infty$.

Lemma 5. *Let $\mathcal{D}$ be a δ-unbiased distribution on $\{-1, 1\}^n$. Suppose that for two vectors $u, w \in \mathbb{R}^n$ and $\theta', \theta'' \in \mathbb{R}$, $\mathbb{E}_{X\sim\mathcal{D}}[(\sigma(\langle w, X\rangle + \theta') - \sigma(\langle u, X\rangle + \theta''))^2] \le \epsilon$, where $\epsilon < \delta e^{-2\|w\|_1 - 2|\theta'| - 6}$. Then $\|w - u\|_\infty \le O(1) \cdot e^{\|w\|_1 + |\theta'|} \cdot \sqrt{\epsilon/\delta}$.*

Lemma 6. *Let $\mathcal{D}$ be a δ-unbiased distribution on $[k]^n$. For $X \sim \mathcal{D}$, let $\tilde{X} \in \{0, 1\}^{n\times k}$ be the one-hot encoded X. Let $u, w \in \mathbb{R}^{n\times k}$ be two matrices satisfying $\sum_a u(i, a) = 0$ and $\sum_a w(i, a) = 0$, for $i \in [n]$. Suppose that for some u, w and $\theta', \theta'' \in \mathbb{R}$, we have $\mathbb{E}_{X\sim\mathcal{D}}[(\sigma(\left\langle w, \tilde{X}\right\rangle + \theta') - \sigma(\left\langle u, \tilde{X}\right\rangle + \theta''))^2] \le \epsilon$, where $\epsilon < \delta e^{-2\|w\|_{\infty,1} - 2|\theta'| - 6}$. Then[5] $\|w - u\|_\infty \le O(1) \cdot e^{\|w\|_{\infty,1} + |\theta'|} \cdot \sqrt{\epsilon/\delta}$.*

The proofs of Lemma 5 and Lemma 6 can be found in [KM17] (see Claim 8.6 and Lemma 4.3 in their paper). We give a slightly different proof of these two lemmas in Appendix E.

3.3 Proof sketches

We provide proof sketches for Theorem 1 and Theorem 2 using the supporting lemmas. The detailed proof can be found in the appendix.

Proof sketch of Theorem 1. Let us focus on the n-th variable (the proof directly extends to other variables). Let $Z \sim \mathcal{D}(A, \theta)$, and $X = [Z_{-n}, 1] = [Z_1, Z_2, \cdots, Z_{n-1}, 1] \in \{-1, 1\}^n$. By Fact 1 and Lemma 1, if $N = O(\lambda^2 \ln(n/\rho)/\gamma^2)$, then $\mathbb{E}_X[(\sigma(\langle w^*, X\rangle) - \sigma(\langle \hat{w}, X\rangle))^2] \le \gamma$ with probability at least $1 - \rho/n$. By Lemma 4 and Lemma 3, Z_{-n} is δ-unbiased with $\delta = e^{-2\lambda}/2$. We can then apply Lemma 5 to show that if $N = O(\lambda^2 \exp(12\lambda) \ln(n/\rho)/\epsilon^4)$, then $\max_{j\in[n]} |A_{nj} - \hat{A}_{nj}| \le \epsilon$ with probability at least $1 - \rho/n$. Theorem 1 then follows by a union bound over all n variables.

Proof sketch of Theorem 2. Let us again focus on the n-th variable since the proof is the same for other variables. As described before, the key step is to show that (14) holds. Now fix a pair of $\alpha \ne \beta \in [k]$, let $N^{\alpha,\beta}$ be the number of samples such that the n-th variable is either α or β. By Lemma 2 and Fact 2, if $N^{\alpha,\beta} = O(\lambda^2 k \ln(n/\rho')/\gamma^2)$, then with probability at least $1 - \rho'$, the matrix $U^{\alpha,\beta} \in \mathbb{R}^{n\times k}$ satisfies $\mathbb{E}_x[(\sigma(\langle w^*, x\rangle) - \sigma(\left\langle U^{\alpha,\beta}, x\right\rangle))^2] \le \gamma$, where $w^* \in \mathbb{R}^{n\times k}$ is defined in (11). By Lemma 6, if $N^{\alpha,\beta} = O(\lambda^2 k^3 \exp(12\lambda) \ln(n/\rho'))/\epsilon^4)$, then with probability at least $1 - \rho'$, $|W_{nj}(\alpha, :) - W_{nj}(\beta, :) - U^{\alpha,\beta}(j, :)| \le \epsilon$, $\forall j \in [n - 1]$. Since $\mathcal{D}(\mathcal{W}, \Theta)$ is δ-unbiased with $\delta = e^{-2\lambda}/k$, in order to have $N^{\alpha,\beta}$ samples for a given (α, β) pair, we need the total number of samples to satisfy $N = O(N^{\alpha,\beta}/\delta)$. Theorem 2 then follows by setting $\rho' = \rho/(nk^2)$ and taking a union bound over all (α, β) pairs and all n variables.

4 Experiments

Learning Ising models. In Figure 1 we construct a diamond-shape graph and show that the incoherence value at Node 1 becomes bigger than 1 (and hence violates the incoherence condition in [RWL10] when we increase the graph size n and edge weight a. We then run 100

[5] For a matrix w, we define $\|w\|_\infty = \max_{ij} |w(i, j)|$. Note that this definition is different from the induced matrix norm.

times of Algorithm 1 and plot the fraction of runs that exactly recovers the underlying graph structure. In each run we generate a different set of samples (sampling is done via exactly computing the distribution). The result shown in Figure 1 is consistent with our analysis and also indicates that our conditions for graph recovery are weaker than those in [RWL10].

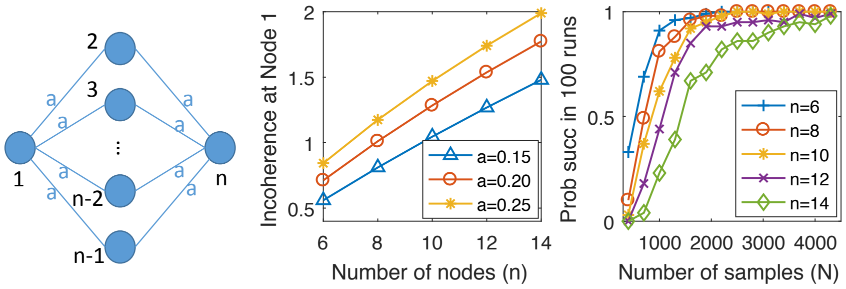

Figure 1: **Left**: The graph structure used in this simulation. It has n nodes and $2(n-2)$ edges. Every edge has the same weight $a > 0$. The mean-field is zero. **Middle**: Incoherence value at Node 1. It violates the incoherence condition in [RWL10] (i.e., becomes larger than 1) when we increase the graph size n and edge weight a. **Right**: We simulate 100 runs of Algorithm 1 for the graph with edge weight $a = 0.2$. The input parameters are $\lambda = 0.2(n-2)$, $\eta = 0.2$.

Learning general pairwise graphical models. We compare our algorithm (Algorithm 2) with the Sparsitron algorithm in [KM17] on a two-dimensional 3-by-3 grid (shown in Figure 2). We experiment three alphabet sizes: $k = 2, 4, 6$. For each value of k, we simulate both algorithms 100 runs, and in each run we generate the W_{ij} matrices with entries ± 0.2. To ensure that each row (as well as each column) of W_{ij} is centered (i.e., zero mean), we will randomly choose W_{ij} between two options: as an example of $k = 2$, $W_{ij} = [0.2, -0.2; -0.2, 0.2]$ or $W_{ij} = [-0.2, 0.2; 0.2, -0.2]$. The mean-field is zero. Sampling is done via exactly computing the distribution. The Sparsitron algorithm requires two sets of samples: 1) to learn a set of candidate weights; 2) to select the best candidate. We use $\max\{200, 0.01 \cdot N\}$ samples for the second part. We plot the estimation error $\max_{ij} \|W_{ij} - \hat{W}_{ij}\|_\infty$ and the fraction of successful runs (i.e., runs that exactly recover the graph) in Figure 3. Compared to the Sparsitron algorithm, our algorithm requires fewer samples for successfully recovering the graphs.

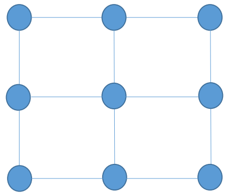

Figure 2: A two-dimensional 3-by-3 grid graph used in the simulation.

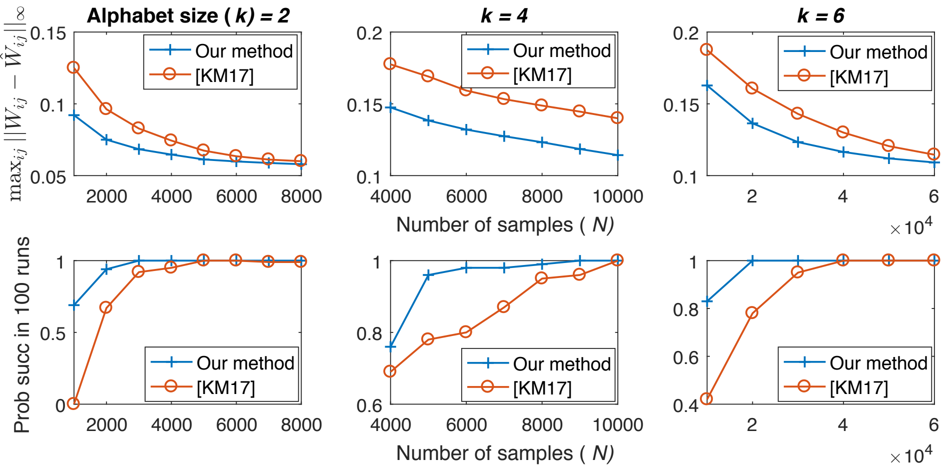

Figure 3: Comparison of our algorithm and the Sparsitron algorithm in [KM17] on a two-dimensional 3-by-3 grid. Top row shows the average of the estimation error $\max_{ij} \|W_{ij} - \hat{W}_{ij}\|_\infty$. Bottom row plots the faction of successful runs (i.e., runs that exactly recover the graph). Each column corresponds to an alphabet size: $k = 2, 4, 6$. Our algorithm needs fewer samples than the Sparsitron algorithm for graph recovery.

References

[AE12] Erik Aurell and Magnus Ekeberg. Inverse ising inference using all the data. *Physical review letters*, 108(9):090201, 2012.

[ANW10] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

[BGd08] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[BM02] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[BM09] José Bento and Andrea Montanari. Which graphical models are difficult to learn? In *Advances in Neural Information Processing Systems*, pages 1303–1311, 2009.

[Bre15] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing (STOC)*, pages 771–782. ACM, 2015.

[BTN13] Ahron Ben-Tal and Arkadi Nemirovski. Lectures on modern convex optimization. https://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf, Fall 2013.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[CLTW10] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 129–136. IEEE, 2010.

[EPL09] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

[HKM17] Linus Hamilton, Frederic Koehler, and Ankur Moitra. Information theoretic properties of markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems*, pages 2463–2472, 2017.

[JRVS11] Ali Jalali, Pradeep Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 378–387, 2011.

[KKB07] Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. An interior-point method for large-scale ℓ_1-regularized logistic regression. *Journal of Machine learning research*, 8(Jul):1519–1555, 2007.

[KM17] Adam R. Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 343–354. IEEE, 2017.

[KSST12] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(Jun):1865–1890, 2012.

[KST09] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pages 793–800, 2009.

[LGK07] Su-In Lee, Varun Ganapathi, and Daphne Koller. Efficient structure learning of markov networks using l_1-regularization. In *Advances in neural Information processing systems*, pages 817–824, 2007.

[LVMC18] Andrey Y Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. Optimal structure and parameter learning of ising models. *Science advances*, 4(3):e1700791, 2018.

[MCK$^+$12] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, The DREAM5 Consortium, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796, 2012.

[RWL10] Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[SW12] 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.

[VMLC16] 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.

[VN49] John Von Neumann. On rings of operators. reduction theory. *Annals of Mathematics*, pages 401–485, 1949.

[YALR12] Eunho Yang, Genevera Allen, Zhandong Liu, and Pradeep K Ravikumar. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366, 2012.

[YL07] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

A Related work on learning Ising models

For the special case of learning Ising models (i.e., binary variables), we compare the sample complexity between the proposed algorithm and the related work in Table 2.

Note that the algorithms in [RWL10, Bre15, VMLC16, LVMC18] are specifically designed for Ising models instead of general pairwise graphical models. That is why they are not presented in Table 1. Our results show that ℓ_1-constrained logistic regression can recover the underlying graph from i.i.d. samples of an Ising model. We make no incoherence assumptions and achieve the state-of-the-art sample complexity.

As mentioned, Ravikumar, Wainwright and Lafferty [RWL10] consider ℓ_1-regularized logistic regression for learning of sparse models in the high-dimensional setting. They require incoherence assumptions that ensure, via conditions on sub-matrices of the Fisher information matrix, that sparse predictors of each node are hard to confuse with a false set. Their analysis obtains significantly better sample complexity compared to what is possible when these extra assumptions are not imposed (see Bento and Montanari [BM09]). Others have also considered ℓ_1-regularization (e.g., [LGK07, YL07, BGd08, JRVS11, YALR12, AE12]) for structure learning of Markov random fields but they all require certain assumptions about the graph and hence their methods do not work for general graphs. The analysis of [RWL10] is of essentially the same convex program as this work (except that we have an additional thresholding procedure). The main difference is that they obtain a better sample guarantee but require significantly more restrictive assumptions.

B Proof of Lemma 1 and Lemma 2

The proof of Lemma 1 relies on the following lemmas. The first lemma is a generalization error bound for any Lipschitz loss of linear functions with bounded ℓ_1-norm.

Lemma 7. *Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_\infty \le X_\infty\}$, and $\mathcal{Y} = \{-1, 1\}$. Let $\ell : \mathbb{R} \to \mathbb{R}$ be a loss function with Lipschitz constant L_ℓ. Define the expected loss $\mathcal{L}(w)$ and the empirical loss $\hat{\mathcal{L}}(w)$ as*

$$\mathcal{L}(w) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell(y \langle w, x\rangle), \quad \hat{\mathcal{L}}(w) = \frac{1}{N}\sum_{i=1}^{N} \ell(y^i \langle w, x^i\rangle), \tag{20}$$

where $\{x^i, y^i\}_{i=1}^N$ are i.i.d. samples from distribution $\mathcal{D}$. Define $\mathcal{W} = \{w \in \mathbb{R}^n : \|w\|_1 \le W_1\}$. Then with probability at least $1-\rho$ over the samples, we have that for all $w \in \mathcal{W}$,

$$\mathcal{L}(w) \le \hat{\mathcal{L}}(w) + 2L_\ell X_\infty W_1 \sqrt{\frac{2\ln(2n)}{N}} + L_\ell X_\infty W_1 \sqrt{\frac{2\ln(2/\rho)}{N}}. \tag{21}$$

Lemma 7 is essentially Theorem 26.15 of [SSBD14] (for the binary classification setup).

Lemma 8. *Let $D_{KL}(a||b) := a\ln(a/b) + (1-a)\ln((1-a)/(1-b))$ denote the KL-divergence between two Bernoulli distributions $(a, 1-a)$, $(b, 1-b)$ with $a, b \in [0, 1]$. Then*

$$(a-b)^2 \le \frac{1}{2} D_{KL}(a||b). \tag{22}$$

Paper	Assumptions	Sample complexity (N)
Information-theoretic lower bound (Thm 1 of [SW12])	1. Model width $\leq \lambda$, and $\lambda \geq 1$ 2. Degree $\leq d$ 3. Minimum edge weight $\geq \eta > 0$ 4. Mean field $= 0$	$\max\{\frac{\ln(n)}{2\eta\tanh(\eta)},$ $\frac{d}{8}\ln(\frac{n}{8d}),$ $\frac{\exp(\lambda)\ln(nd/4-1)}{4\eta d\exp(\eta)}\}$
ℓ_1-regularized logistic regression [RWL10]	Q^* is the Fisher information matrix, S is set of neighbors of a given variable. 1. Dependency: $\exists\ C_{\min} > 0$ such that eigenvalues of $Q^*_{SS} \geq C_{\min}$ 2. Incoherence: $\exists\ \alpha \in (0, 1]$ such that $\lVert Q^*_{S^cS}(Q^*_{SS})^{-1}\rVert_\infty \leq 1-\alpha$ 3. Regularization parameter: $\lambda_N \geq \frac{16(2-\alpha)}{\alpha}\sqrt{\frac{\ln(n)}{N}}$ 4. Minimum edge weight $\geq 10\sqrt{d}\lambda_N/C_{\min}$ 5. Mean field $= 0$ 6. Probability of success $\geq 1-2e^{-O(\lambda_N^2 N)}$	$O(d^3\ln(n))$
Greedy algorithm [Bre15]	1. Model width $\leq \lambda$ 2. Degree $\leq d$ 3. Minimum edge weight $\geq \eta > 0$ 4. Probability of success $\geq 1-\rho$	$O(\exp(\frac{\exp(O(d\lambda))}{\eta^{O(1)}})\ln(\frac{n}{\rho}))$
Interaction Screening [VMLC16]	1. Model width $\leq \lambda$ 2. Degree $\leq d$ 3. Minimum edge weight $\geq \eta > 0$ 4. Regularization parameter $= 4\sqrt{\frac{\ln(3n^2/\rho)}{N}}$ 5. Probability of success $\geq 1-\rho$	$O(\max\{d, \frac{1}{\eta^2}\}$ $d^3\exp(6\lambda)\ln(\frac{n}{\rho}))$
ℓ_1-regularized logistic regression [LVMC18]	1. Model width $\leq \lambda$ 2. Degree $\leq d$ 3. Minimum edge weight $\geq \eta > 0$ 4. Regularization parameter $O(\sqrt{\frac{\ln(n^2/\rho)}{N}})$ 5. Probability of success $\geq 1-\rho$	$O(\max\{d, \frac{1}{\eta^2}\}$ $d^3\exp(8\lambda)\ln(\frac{n}{\rho}))$
Sparsitron [KM17]	1. Model width $\leq \lambda$ 2. Minimum edge weight $\geq \eta > 0$ 3. Probability of success $\geq 1-\rho$	$O(\frac{\lambda^2\exp(12\lambda)}{\eta^4}\ln(\frac{n}{\rho\eta}))$
ℓ_1-constrained logistic regression [this paper]	1. Model width $\leq \lambda$ 2. Minimum edge weight $\geq \eta > 0$ 3. Probability of success $\geq 1-\rho$	$O(\frac{\lambda^2\exp(12\lambda)}{\eta^4}\ln(\frac{n}{\rho}))$

Table 2: Comparison of the sample complexity required for graph recovery of an Ising model. The second column lists the assumptions in their analysis. Given λ and η, degree $d \leq \lambda/\eta$.

Lemma 8 is simply the Pinsker's inequality applied to the binary distributions.

Lemma 9. *Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \{-1, 1\}$ where for $(X, Y) \sim \mathcal{D}$, $\mathbb{P}[Y = 1|X = x] = \sigma(\langle w^*, x\rangle)$. Let $\mathcal{L}(w)$ be the expected logistic loss:*

$$\mathcal{L}(w) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ln(1+e^{-y\langle w,x\rangle}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[-\frac{y+1}{2}\ln(\sigma(\langle w, x\rangle))-\frac{1-y}{2}\ln(1-\sigma(\langle w, x\rangle))]. \tag{23}$$

Then for any w, we have

$$\mathcal{L}(w) - \mathcal{L}(w^*) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[D_{KL}(\sigma(\langle w^*, x\rangle)||\sigma(\langle w, x\rangle))], \tag{24}$$

where $D_{KL}(a||b) := a\ln(a/b) + (1 - a)\ln((1 - a)/(1 - b))$ denotes the KL-divergence between two Bernoulli distributions $(a, 1 - a)$, $(b, 1 - b)$ with $a, b \in [0, 1]$, and $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function.

Proof. Simply plugging in the definition of the expected logistic loss $\mathcal{L}(\cdot)$ gives

$$\begin{aligned}
\mathcal{L}(w) - \mathcal{L}(w^*) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[-\frac{y+1}{2}\ln(\sigma(\langle w, x\rangle)) - \frac{1-y}{2}\ln(1-\sigma(\langle w, x\rangle))] \\
&\quad + \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\frac{y+1}{2}\ln(\sigma(\langle w^*, x\rangle)) + \frac{1-y}{2}\ln(1-\sigma(\langle w^*, x\rangle))] \\
&= \mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y|x}[-\frac{y+1}{2}\ln(\sigma(\langle w, x\rangle)) - \frac{1-y}{2}\ln(1-\sigma(\langle w, x\rangle))] \\
&\quad + \mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{y|x}[\frac{y+1}{2}\ln(\sigma(\langle w^*, x\rangle)) + \frac{1-y}{2}\ln(1-\sigma(\langle w^*, x\rangle))] \\
&\overset{(a)}{=} \mathop{\mathbb{E}}_{x}[-\sigma(\langle w^*, x\rangle)\ln(\sigma(\langle w, x\rangle)) - (1-\sigma(\langle w^*, x\rangle))\ln(1-\sigma(\langle w, x\rangle))] \\
&\quad + \mathop{\mathbb{E}}_{x}[\sigma(\langle w^*, x\rangle)\ln(\sigma(\langle w^*, x\rangle)) + (1-\sigma(\langle w^*, x\rangle))\ln(1-\sigma(\langle w^*, x\rangle))] \\
&= \mathop{\mathbb{E}}_{x}\left[\sigma(\langle w^*, x\rangle)\ln\left(\frac{\sigma(\langle w^*, x\rangle)}{\sigma(\langle w, x\rangle)}\right) + (1-\sigma(\langle w^*, x\rangle))\ln\left(\frac{1-\sigma(\langle w^*, x\rangle)}{1-\sigma(\langle w, x\rangle)}\right)\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[D_{KL}(\sigma(\langle w^*, x\rangle)||\sigma(\langle w, x\rangle))],
\end{aligned}$$

where (a) follows from the fact that

$$E_{y|x}[y] = 1 \cdot \mathbb{P}[y = 1|x] + (-1) \cdot \mathbb{P}[y = -1|x] = 2\sigma(\langle w^*, x\rangle) - 1.$$

□

We are now ready to prove Lemma 1 (which is restated below):

Lemma. *Let $\mathcal{D}$ be a distribution on $\{-1, 1\}^n \times \{-1, 1\}$ where for $(X, Y) \sim \mathcal{D}$, $\mathbb{P}[Y = 1|X = x] = \sigma(\langle w^*, x\rangle)$. We assume that $\|w^*\|_1 \le 2\lambda$ for a known $\lambda \ge 0$. Given N i.i.d. samples $\{(x^i, y^i)\}_{i=1}^N$, let $\hat{w}$ be any minimizer of the following ℓ_1-constrained logistic regression problem:*

$$\hat{w} \in \arg\min_{w\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N}\ln(1 + e^{-y^i\langle w,x^i\rangle}) \quad \text{s.t. } \|w\|_1 \le 2\lambda. \tag{25}$$

Given $\rho \in (0, 1)$ and $\epsilon > 0$, suppose that $N = O(\lambda^2(\ln(n/\rho))/\epsilon^2)$, then with probability at least $1 - \rho$ over the samples, we have that $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(\langle w^*, x\rangle) - \sigma(\langle \hat{w}, x\rangle))^2] \le \epsilon$.

Proof. We first apply Lemma 7 to the setup of Lemma 1. The loss function $\ell(z) = \ln(1 + e^{-z})$ defined above has Lipschitz constant $L_\ell = 1$. The input sample $x \in \{-1, 1\}^n$ satisfies $\|x\|_\infty \leq 1$. Let $\mathcal{W} = \{w \in \mathbb{R}^{n \times k} : \|w\|_1 \leq 2\lambda\}$. According to Lemma 7, with probability at least $1 - \rho/2$ over the draw of the training set, we have that for all $w \in \mathcal{W}$,

$$\mathcal{L}(w) - \hat{\mathcal{L}}(w) \leq 4\lambda\sqrt{\frac{2\ln(2n)}{N}} + 2\lambda\sqrt{\frac{2\ln(4/\rho)}{N}}. \tag{26}$$

where $\mathcal{L}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \ln(1 + e^{-y\langle w,x\rangle})$ and $\hat{\mathcal{L}}(w) = \sum_{i=1}^N \ln(1 + e^{-y^i\langle w,x^i\rangle})/N$ are the expected loss and empirical loss.

Let $N = C \cdot \lambda^2 \ln(8n/\rho)/\epsilon^2$ for a global constant C, then (26) implies that with probability at least $1 - \rho/2$,

$$\mathcal{L}(w) \leq \hat{\mathcal{L}}(w) + \epsilon, \text{ for all } w \in \mathcal{W}. \tag{27}$$

We next prove a concentration result for $\hat{\mathcal{L}}(w^*)$. Here w^* is the true regression vector and is assumed to be fixed. Since $\ln(1 + e^{-y\langle w^*,x\rangle})$ is bounded for $x \in \mathcal{X}$ and $w^* \in \mathcal{W}$, Hoeffding's inequality gives that $\mathbb{P}[\hat{\mathcal{L}}(w^*) - \mathcal{L}(w^*) \geq t] \leq e^{-2Nt^2/(4\lambda)^2}$. Let $N = C' \cdot \lambda^2 \ln(2/\rho)/\epsilon^2$ for a global constant C', then with probability at least $1 - \rho/2$ over the samples,

$$\hat{\mathcal{L}}(w^*) \leq \mathcal{L}(w^*) + \epsilon. \tag{28}$$

Then the following holds with probability at least $1 - \rho$:

$$\mathcal{L}(\hat{w}) \overset{(a)}{\leq} \hat{\mathcal{L}}(\hat{w}) + \epsilon \overset{(b)}{\leq} \hat{\mathcal{L}}(w^*) + \epsilon \overset{(c)}{\leq} \mathcal{L}(w^*) + 2\epsilon, \tag{29}$$

where (a) follows from (27), (b) follows from the fact $\hat{w}$ is the minimizer of $\hat{\mathcal{L}}(w)$, and (c) follows from (28).

So far we have shown that $\mathcal{L}(\hat{w}) - \mathcal{L}(w^*) \leq 2\epsilon$ with probability at least $1 - \rho$. The last step is to lower bound $\mathcal{L}(\hat{w}) - \mathcal{L}(w^*)$ by $\mathbb{E}_{(x,y)\sim\mathcal{D}}(\sigma(\langle w^*, x\rangle) - \sigma(\langle w, x\rangle))^2$ using Lemma 8 and Lemma 9.

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}(\sigma(\langle w^*, x\rangle) - \sigma(\langle w, x\rangle))^2 &\overset{(d)}{\leq} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} D_{KL}(\sigma(\langle w^*, x\rangle)||\sigma(\langle w, x\rangle))/2 \\
&\overset{(e)}{=} (\mathcal{L}(\hat{w}) - \mathcal{L}(w^*))/2 \\
&\overset{(f)}{\leq} \epsilon,
\end{aligned}$$

where (d) follows from Lemma 8, (e) follows from Lemma 9, and (f) follows from (29). Therefore, we have that $\mathbb{E}_{(x,y)\sim\mathcal{D}}(\sigma(\langle w^*, x\rangle) - \sigma(\langle w, x\rangle))^2 \leq \epsilon$ with probability at least $1 - \rho$, if the number of samples satisfies $N = O(\lambda^2 \ln(n/\rho)/\epsilon^2)$. □

The proof of Lemma 2 is identical to the proof of Lemma 1, except that it relies on the following generalization error bound for Lipschitz loss functions with bounded $\ell_{2,1}$-norm.

Lemma 10. *Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{x \in \mathbb{R}^{n \times k} : \|x\|_{2,\infty} \leq X_{2,\infty}\}$, and $\mathcal{Y} = \{-1, 1\}$. Let $\ell : \mathbb{R} \to \mathbb{R}$ be a loss function with Lipschitz constant L_ℓ. Define the expected loss $\mathcal{L}(w)$ and the empirical loss $\hat{\mathcal{L}}(w)$ as*

$$\mathcal{L}(w) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell(y\langle w, x\rangle), \quad \hat{\mathcal{L}}(w) = \frac{1}{N}\sum_{i=1}^N \ell(y^i\langle w, x^i\rangle), \tag{30}$$

where $\{x^i, y^i\}_{i=1}^N$ are i.i.d. samples from distribution $\mathcal{D}$. Define $\mathcal{W} = \{w \in \mathbb{R}^{n\times k} : \|w\|_{2,1} \leq W_{2,1}\}$. Then with probability at least $1-\rho$ over the draw of N samples, we have that for all $w \in \mathcal{W}$,

$$\mathcal{L}(w) \leq \hat{\mathcal{L}}(w) + 2L_\ell X_{2,\infty} W_{2,1}\sqrt{\frac{6\ln(n)}{N}} + L_\ell X_{2,\infty} W_{2,1}\sqrt{\frac{2\ln(2/\rho)}{N}}. \tag{31}$$

Lemma 10 can be readily derived from the existing results. First, notice that the dual norm of $\|\cdot\|_{2,1}$ is $\|\cdot\|_{2,\infty}$. Using Corollary 14 in [KSST12], Theorem 1 in [KST09], and the fact that $\|w\|_{2,q} \leq \|w\|_{2,1}$ for $q \geq 1$, we conclude that the Rademacher complexity of the function class $\mathcal{F} := \{x \to \langle w, x\rangle : \|w\|_{2,1} \leq W_{2,1}\}$ is at most $X_{2,\infty}W_{2,1}\sqrt{6\ln(n)/N}$. We can then obtain the standard Rademacher-based generalization bound (see, e.g., [BM02] and Theorem 26.5 in [SSBD14]) for bounded Lipschitz loss functions.

We omit the proof of Lemma 2 since it is the same as that of Lemma 1.

C Proof of Lemma 3

Lemma 3 is restated below.

Lemma. *Let $\mathcal{D}$ be a δ-unbiased distribution on S^n, where S is the alphabet set. For $X \sim \mathcal{D}$, any $i \in [n]$, the distribution of X_{-i} is also δ-unbiased.*

Proof. For any $j \neq i \in [n]$, any $a \in S$, and any $x \in S^{n-2}$, we have

$$\begin{aligned}\mathbb{P}[X_j = a | X_{[n]\setminus\{i,j\}} = x] &= \sum_{b\in S}\mathbb{P}[X_j = a, X_i = b | X_{[n]\setminus\{i,j\}} = x]\\ &= \sum_{b\in S}\mathbb{P}[X_i = b | X_{[n]\setminus\{i,j\}} = x]\cdot\mathbb{P}[X_j = a | X_i = b, X_{[n]\setminus\{i,j\}} = x]\\ &\overset{(a)}{\leq} \delta\sum_{b\in S}\mathbb{P}[X_i = b | X_{[n]\setminus\{i,j\}} = x]\\ &= \delta,\end{aligned} \tag{32}$$

where (a) follows from the fact that $X \sim \mathcal{D}$ and $\mathcal{D}$ is a δ-unbiased distribution. Since (32) holds for any $j \neq i \in [n]$, any $a \in S$, and any $x \in S^{n-2}$, by definition, the distribution of X_{-i} is δ-unbiased. □

D Proof of Lemma 4

The lemma is restated below, followed by its proof.

Lemma. *Let $\mathcal{D}(\mathcal{W}, \Theta)$ be a pairwise graphical model distribution with alphabet size k and width $\lambda(\mathcal{W}, \Theta)$. Then $\mathcal{D}(\mathcal{W}, \Theta)$ is δ-unbiased with $\delta = e^{-2\lambda(\mathcal{W},\Theta)}/k$. Specifically, an Ising model distribution $\mathcal{D}(A, \theta)$ is $e^{-2\lambda(A,\theta)}/2$-unbiased.*

Proof. Let $X \sim \mathcal{D}(\mathcal{W}, \Theta)$, and assume that $X \in [k]^n$. For any $i \in [n]$, any $a \in [k]$, and any $x \in [k]^{n-1}$, we have

$$\begin{aligned}
\mathbb{P}[X_i = a | X_{-i} = x] &= \frac{\exp(\sum_{j \neq i} W_{ij}(a, x_j) + \theta_i(a))}{\sum_{b \in [k]} \exp(\sum_{j \neq i} W_{ij}(b, x_j) + \theta_i(b))} \\
&= \frac{1}{\sum_{b \in [k]} \exp(\sum_{j \neq i}(W_{ij}(b, x_j) - W_{ij}(a, x_j)) + \theta_i(b) - \theta_i(a))} \\
&\overset{(a)}{\geq} \frac{1}{k \cdot \exp(2\lambda(\mathcal{W}, \Theta))} \\
&= e^{-2\lambda(\mathcal{W}, \Theta)}/k, \qquad (33)
\end{aligned}$$

where (a) follows from the definition of model width. The lemma then follows (Ising model corresponds to the special case of $k = 2$). □

E Proof of Lemma 5 and Lemma 6

The proof relies on the following basic property of the sigmoid function (see Claim 4.2 of [KM17]):

$$|\sigma(a) - \sigma(b)| \geq e^{-|a|-3} \cdot \min(1, |a - b|), \quad \forall a, b \in \mathbb{R}. \qquad (34)$$

We first prove Lemma 5 (which is restated below).

Lemma. *Let $\mathcal{D}$ be a δ-unbiased distribution on $\{-1, 1\}^n$. Suppose that for two vectors $u, w \in \mathbb{R}^n$ and $\theta', \theta'' \in \mathbb{R}$, $\mathbb{E}_{X \sim \mathcal{D}}[(\sigma(\langle w, X \rangle + \theta') - \sigma(\langle u, X \rangle + \theta''))^2] \leq \epsilon$, where $\epsilon < \delta e^{-2\|w\|_1 - 2|\theta'| - 6}$. Then $\|w - u\|_\infty \leq O(1) \cdot e^{\|w\|_1 + |\theta'|} \cdot \sqrt{\epsilon/\delta}$.*

Proof. For any $i \in [n]$, any $X \in \{-1, 1\}^n$, let $X_i \in \{-1, 1\}$ be the i-th variable and $X_{-i} \in \{-1, 1\}^{n-1}$ be the $[n]\backslash\{i\}$ variables. Let $X^{i,+} \in \{-1, 1\}^n$ (respectively $X^{i,-}$) be the vector obtained from X by setting $X_i = 1$ (respectively $X_i = -1$). Then we have

$$\begin{aligned}
\epsilon &\geq \mathop{\mathbb{E}}_{X \sim \mathcal{D}}[(\sigma(\langle w, X \rangle + \theta') - \sigma(\langle u, X \rangle + \theta''))^2] \\
&= \mathop{\mathbb{E}}_{X_{-i}}\left[\mathop{\mathbb{E}}_{X_i | X_{-i}} (\sigma(\langle w, X \rangle + \theta') - \sigma(\langle u, X \rangle + \theta''))^2\right] \\
&= \mathop{\mathbb{E}}_{X_{-i}}[(\sigma(\langle w, X^{i,+} \rangle + \theta') - \sigma(\langle u, X^{i,+} \rangle + \theta''))^2 \cdot \mathbb{P}[X_i = 1 | X_{-i}] \\
&\qquad + (\sigma(\langle w, X^{i,-} \rangle + \theta') - \sigma(\langle u, X^{i,-} \rangle + \theta''))^2 \cdot \mathbb{P}[X_i = -1 | X_{-i}]] \\
&\overset{(a)}{\geq} \delta \cdot \mathop{\mathbb{E}}_{X_{-i}}[(\sigma(\langle w, X^{i,+} \rangle + \theta') - \sigma(\langle u, X^{i,+} \rangle + \theta''))^2 \\
&\qquad + (\sigma(\langle w, X^{i,-} \rangle + \theta') - \sigma(\langle u, X^{i,-} \rangle + \theta''))^2] \\
&\overset{(b)}{\geq} \delta e^{-2\|w\|_1 - 2|\theta'| - 6} \cdot \mathop{\mathbb{E}}_{X_{-i}}[\min(1, ((\langle w, X^{i,+} \rangle + \theta') - (\langle u, X^{i,+} \rangle + \theta''))^2) \\
&\qquad + \min(1, ((\langle w, X^{i,-} \rangle + \theta') - (\langle u, X^{i,-} \rangle + \theta''))^2)] \\
&\overset{(c)}{\geq} \delta e^{-2\|w\|_1 - 2|\theta'| - 6} \cdot \mathop{\mathbb{E}}_{X_{-i}} \min(1, (2w_i - 2u_i)^2/2) \\
&\overset{(d)}{=} \delta e^{-2\|w\|_1 - 2|\theta'| - 6} \cdot \min(1, 2(w_i - u_i)^2). \qquad (35)
\end{aligned}$$

Here (a) follows from the fact that $\mathcal{D}$ is a δ-unbiased distribution, which implies that $\mathbb{P}[X_i = 1|X_{-i}] \geq \delta$ and $\mathbb{P}[X_i = -1|X_{-i}] \geq \delta$. Inequality (b) is obtained by substituting (34). Inequality (c) uses the following fact

$$\min(1, a^2) + \min(1, b^2) \geq \min(1, (a-b)^2/2), \forall a, b \in \mathbb{R}. \tag{36}$$

To see why (36) holds, note that if both $|a|, |b| \leq 1$, then (36) is true since $a^2 + b^2 \geq (a-b)^2/2$. Otherwise, (36) is true because the left-hand side is at least 1 while the right-hand side is at most 1. The last equality (d) follows from that X_{-i} is independent of $\min(1, 2(w_i - u_i)^2)$.

Since $\epsilon < \delta e^{-2\|w\|_1 - 2|\theta'| - 6}$, (35) implies that $|w_i - u_i| \leq O(1) \cdot e^{\|w\|_1 + |\theta'|} \cdot \sqrt{\epsilon/\delta}$. Because (35) holds for any $i \in [n]$, we have that $\|w - u\|_\infty \leq O(1) \cdot e^{\|w\|_1 + |\theta'|} \cdot \sqrt{\epsilon/\delta}$. □

We now prove Lemma 6 (which is restated below).

Lemma. *Let $\mathcal{D}$ be a δ-unbiased distribution on $[k]^n$. For $X \sim \mathcal{D}$, let $\tilde{X} \in \{0,1\}^{n \times k}$ be the one-hot encoded X. Let $u, w \in \mathbb{R}^{n \times k}$ be two matrices satisfying $\sum_j u(i,j) = 0$ and $\sum_j w(i,j) = 0$, for $i \in [n]$. Suppose that for some u, w and $\theta', \theta'' \in \mathbb{R}$, we have $\mathbb{E}_{X \sim \mathcal{D}}[(\sigma(\langle w, \tilde{X} \rangle + \theta') - \sigma(\langle u, \tilde{X} \rangle + \theta''))^2] \leq \epsilon$, where $\epsilon < \delta e^{-2\|w\|_{\infty,1} - 2|\theta'| - 6}$. Then $\|w - u\|_\infty \leq O(1) \cdot e^{\|w\|_{\infty,1} + |\theta'|} \cdot \sqrt{\epsilon/\delta}$.*

Proof. Fix an $i \in [n]$ and $a \neq b \in [k]$. Let $X^{i,a} \in [k]^n$ (respectively $X^{i,b}$) be the vector obtained from X by setting $X_i = a$ (respectively $X_i = b$). Let $\tilde{X}^{i,a} \in \{0,1\}^{n \times k}$ be the one-hot encoding of $X^{i,a} \in [k]^n$. Then we have

$$\begin{aligned}
\epsilon &\geq \mathop{\mathbb{E}}_{X \sim \mathcal{D}}[(\sigma(\langle w, \tilde{X} \rangle + \theta') - \sigma(\langle u, \tilde{X} \rangle + \theta''))^2] \\
&= \mathop{\mathbb{E}}_{X_{-i}} \left[\mathop{\mathbb{E}}_{X_i | X_{-i}} (\sigma(\langle w, \tilde{X} \rangle + \theta') - \sigma(\langle u, \tilde{X} \rangle + \theta''))^2 \right] \\
&\geq \mathop{\mathbb{E}}_{X_{-i}} [(\sigma(\langle w, \tilde{X}^{i,a} \rangle + \theta') - \sigma(\langle u, \tilde{X}^{i,a} \rangle + \theta''))^2 \cdot \mathbb{P}[X_i = a | X_{-i}] \\
&\qquad + (\sigma(\langle w, \tilde{X}^{i,b} \rangle + \theta') - \sigma(\langle u, \tilde{X}^{i,b} \rangle + \theta''))^2 \cdot \mathbb{P}[X_i = b | X_{-i}]] \\
&\overset{(a)}{\geq} \delta e^{-2\|w\|_{\infty,1} - 2|\theta'| - 6} \cdot \mathop{\mathbb{E}}_{X_{-i}} [\min(1, ((\langle w, \tilde{X}^{i,a} \rangle + \theta') - (\langle u, \tilde{X}^{i,a} \rangle + \theta''))^2) \\
&\qquad\qquad + \min(1, ((\langle w, \tilde{X}^{i,b} \rangle + \theta') - (\langle u, \tilde{X}^{i,b} \rangle + \theta''))^2)] \\
&\overset{(b)}{\geq} \delta e^{-2\|w\|_{\infty,1} - 2|\theta'| - 6} \cdot \mathop{\mathbb{E}}_{X_{-i}} \min(1, ((w(i,a) - w(i,b)) - (u(i,a) - u(i,b)))^2/2) \\
&= \delta e^{-2\|w\|_{\infty,1} - 2|\theta'| - 6} \min(1, ((w(i,a) - w(i,b)) - (u(i,a) - u(i,b)))^2/2)
\end{aligned} \tag{37}$$

Here (a) follows from that $\mathcal{D}$ is a δ-unbiased distribution and (34). Inequality (b) follows from (36). Because $\epsilon < \delta e^{-2\|w\|_{\infty,1} - 2|\theta'| - 6}$, (37) implies that

$$(w(i,a) - w(i,b)) - (u(i,a) - u(i,b)) \leq O(1) \cdot e^{\|w\|_{\infty,1} + |\theta'|} \cdot \sqrt{\epsilon/\delta}. \tag{38}$$

$$(u(i,a) - u(i,b)) - (w(i,a) - w(i,b)) \leq O(1) \cdot e^{\|w\|_{\infty,1} + |\theta'|} \cdot \sqrt{\epsilon/\delta}. \tag{39}$$

Since (38) and (39) hold for any $a \neq b \in [k]$, we can sum over $b \in [k]$ and use the fact that $\sum_j u(i,j) = 0$ and $\sum_j w(i,j) = 0$ to get

$$w(i,a) - u(i,a) = \frac{1}{k}\sum_b (w(i,a) - w(i,b)) - (u(i,a) - u(i,b)) \leq O(1) \cdot e^{\|w\|_{\infty,1} + |\theta'|} \cdot \sqrt{\epsilon/\delta}.$$

$$u(i,a) - w(i,a) = \frac{1}{k}\sum_b (u(i,a) - u(i,b)) - (w(i,a) - w(i,b)) \leq O(1) \cdot e^{\|w\|_{\infty,1} + |\theta'|} \cdot \sqrt{\epsilon/\delta}.$$

Therefore, we have $|w(i,a) - u(i,a)| \leq O(1) \cdot e^{\|w\|_{\infty,1} + |\theta'|} \cdot \sqrt{\epsilon/\delta}$, for any $i \in [n]$ and $a \in [k]$. □

F Proof of Theorem 1

We first restate Theorem 1 and then give the proof.

Theorem. *Let $\mathcal{D}(A,\theta)$ be an unknown n-variable Ising model distribution with dependency graph G. Suppose that the $\mathcal{D}(A,\theta)$ has width $\lambda(A,\theta) \leq \lambda$. Given $\rho \in (0,1)$ and $\epsilon > 0$, if the number of i.i.d. samples satisfies $N = O(\lambda^2 \exp(12\lambda)\ln(n/\rho)/\epsilon^4)$, then with probability at least $1-\rho$, Algorithm 1 produces $\hat{A}$ that satisfies*

$$\max_{i,j \in [n]} |A_{ij} - \hat{A}_{ij}| \leq \epsilon. \tag{40}$$

Proof. For simplicity, we focus on the n-th variable, and will show that Algorithm 1 is able to recover the n-th row of the true weight matrix A. Specifically, we will show that if the number samples satisfies $N = O(\lambda^2 \exp(O(\lambda))\ln(n/\rho)/\epsilon^4)$, then with probability as least $1 - \rho/n$,

$$\max_{j \in [n]} |A_{nj} - \hat{A}_{nj}| \leq \epsilon. \tag{41}$$

The proof directly extends to other variables. We can then use a union bound to conclude that with probability as least $1-\rho$, $\max_{i,j\in[n]} |A_{ij} - \hat{A}_{ij}| \leq \epsilon$.

To show that Eq. (41) holds, let $Z \sim \mathcal{D}(A,\theta)$, $X = (Z_1, Z_2, \cdots, Z_{n-1}, 1) \in \{-1,1\}^n$, $Y = Z_n \in \{-1,1\}$, by Fact 1, we have that

$$\mathbb{P}[Y = 1 | X = x] = \sigma(\langle w^*, x\rangle), \quad \text{where } w^* = 2(A_{n1}, A_{n2}, \cdots, A_{n(n-1)}, \theta_n) \in \mathbb{R}^n. \tag{42}$$

In Algorithm 1, we form N samples $\{(x^i, y^i)\}_{i=1}^N$ that satisfy Eq. (42). Furthermore, $\|w^*\|_1 \leq 2\lambda(A,\theta) \leq 2\lambda$, by the definition of model width. Then an ℓ_1-constrained logistic regression is solved and the output is $\hat{w} \in \mathbb{R}^n$.

By Lemma 1, if the number of samples satisfies $N = O(\lambda^2 \ln(n/\rho)/\gamma^2)$, then with probability at least $1-\rho/n$, we have

$$\mathop{\mathbb{E}}_X[(\sigma(\langle w^*, X\rangle) - \sigma(\langle \hat{w}, X\rangle))^2] \leq \gamma, \tag{43}$$

where $X = (Z_{-n}, 1) = (Z_1, Z_2, \cdots, Z_{n-1}, 1) \in \{-1,1\}^n$.

By Lemma 4, $Z_{-n} \in \{-1,1\}^{n-1}$ is δ-unbiased (Definition 3) with $\delta = e^{-2\lambda}/2$. Applying Lemma 5 to Eq. (43) gives

$$\|w^*_{1:(n-1)} - \hat{w}_{1:(n-1)}\|_\infty \leq O(1) \cdot e^{2\lambda} \cdot \sqrt{\gamma/\delta} \tag{44}$$

for $\gamma < C_1 \delta e^{-4\lambda}$ for some constant $C_1 > 0$. Given $\epsilon \in (0,1)$, we now set $\gamma = C_2 \delta e^{-4\lambda} \epsilon^2$ for some constant $C_2 > 0$. Note that $w^*_{1:(n-1)} = 2(A_{n1}, \cdots, A_{n(n-1)})$ and $\hat{w}_{1:(n-1)} = 2(\hat{A}_{n1}, \cdots, \hat{A}_{n(n-1)})$. Eq. (44) implies that

$$\max_{j \in [n]} |A_{nj} - \hat{A}_{nj}| \leq \epsilon. \tag{45}$$

The number of samples needed is $N = O(\lambda^2 \ln(n/\rho)/\gamma^2) = O(\lambda^2 e^{12\lambda} \ln(n/\rho)/\epsilon^4)$.

We have shown that Eq. (41) holds with probability at least $1 - \rho/n$. Using a union bound over all n variables gives that with probability as least $1 - \rho$, $\max_{i,j \in [n]} |A_{ij} - \hat{A}_{ij}| \leq \epsilon$. □

G Proof of Theorem 2

Theorem 2 is restated below, followed by its proof.

Theorem. *Let $\mathcal{D}(\mathcal{W}, \Theta)$ be an n-variable pairwise graphical model distribution with width $\lambda(\mathcal{W}, \Theta) \leq \lambda$ and alphabet size k. Given $\rho \in (0,1)$ and $\epsilon > 0$, if the number of i.i.d. samples satisfies $N = O(\lambda^2 k^4 \exp(14\lambda) \ln(nk/\rho)/\epsilon^4)$, then with probability at least $1 - \rho$, Algorithm 2 produces $\hat{W}_{ij} \in \mathbb{R}^{k \times k}$ that satisfies*

$$|W_{ij}(a,b) - \hat{W}_{ij}(a,b)| \leq \epsilon, \quad \forall i \neq j \in [n], \ \forall a, b \in [k]. \tag{46}$$

Proof. For simplicity, let us focus on the n-th variable (i.e., set $i = n$ inside the first "for" loop of Algorithm 2). The proof directly applies to other variables. We will prove the following result: if the number of samples $N = O(\lambda^2 k^4 \exp(O(\lambda)) \ln(nk/\rho)/\epsilon^4)$, then with probability at least $1 - \rho/n$, the $U^{\alpha,\beta} \in \mathbb{R}^{n \times k}$ matrices produced by Algorithm 2 satisfies

$$|W_{nj}(\alpha, :) - W_{nj}(\beta, :) - U^{\alpha,\beta}(j, :)| \leq \epsilon, \quad \forall j \in [n-1], \ \forall \alpha, \beta \in [k]. \tag{47}$$

Suppose that (47) holds, then summing over $\beta \in [k]$ and using the fact that $\sum_\beta W_{nj}(\beta, :) = 0$ gives

$$|W_{nj}(\alpha, :) - \frac{1}{k} \sum_{\beta \in [k]} U^{\alpha,\beta}(j, :)| \leq \epsilon, \quad \forall j \in [n-1], \ \forall \alpha \in [k]. \tag{48}$$

Since $\hat{W}_{ij}(\alpha, :) = \sum_{\beta \in [k]} U^{\alpha,\beta}(j, :)/k$, Theorem 2 then follows by taking a union bound over the n variables.

The only thing left is to prove (47). Now fix a pair of $\alpha, \beta \in [k]$, let $N^{\alpha,\beta}$ be the number of samples such that the n-th variable is either α or β. By Lemma 2 and Fact 2, if $N^{\alpha,\beta} = O(\lambda^2 k \ln(n/\rho')/\gamma^2)$, then with probability at least $1 - \rho'$, the minimizer of the $\ell_{2,1}$ constrained logistic regression $w^{\alpha,\beta} \in \mathbb{R}^{n \times k}$ satisfies

$$\mathop{\mathbb{E}}_X [(\sigma(\langle w^*, X \rangle) - \sigma(\left\langle w^{\alpha,\beta}, X \right\rangle))^2] \leq \gamma, \tag{49}$$

where the random variable $X \in \{0,1\}^{n \times k}$ is the one-hot encoding of vector $(Z_{-n}, 1) \in [k]^n$ for $Z \sim \mathcal{D}(\mathcal{W}, \Theta)$, and $w^* \in \mathbb{R}^{n \times k}$ satisfies

$$w^*(j, :) = W_{nj}(\alpha, :) - W_{nj}(\beta, :), \ \forall j \in [n-1];$$
$$w^*(n, :) = [\theta_n(\alpha) - \theta_n(\beta), 0, \cdots, 0].$$

Recall that $U^{\alpha,\beta} \in \mathbb{R}^{n\times k}$ is formed by centering the first $n-1$ rows of $w^{\alpha,\beta}$. Because each row of X is a standard basis vector (i.e., all 0's except a single 1), we have $\left\langle U^{\alpha,\beta}, X\right\rangle = \left\langle w^{\alpha,\beta}, X\right\rangle$. Hence, (49) implies that

$$\mathop{\mathbb{E}}_{X}[(\sigma(\langle w^*, X\rangle) - \sigma(\left\langle U^{\alpha,\beta}, X\right\rangle))^2] \leq \gamma. \tag{50}$$

By Lemma 4 and Lemma 3, for $Z \sim \mathcal{D}(\mathcal{W}, \Theta)$, Z_{-n} is δ-unbiased with $\delta = e^{-2\lambda}/k$. By Lemma 6 and (50), if $N^{\alpha,\beta} = O(\lambda^2 k^3 \exp(12\lambda)\ln(n/\rho'))/\epsilon^4)$, then with probability at least $1-\rho'$,

$$|W_{nj}(\alpha,:) - W_{nj}(\beta,:) - U^{\alpha,\beta}(j,:)| \leq \epsilon, \ \forall j \in [n-1]. \tag{51}$$

So far we have proved that (47) holds for a fixed (α,β) pair. This requires that $N^{\alpha,\beta} = O(\lambda^2 k^3 \exp(12\lambda)\ln(n/\rho'))/\epsilon^4)$. Recall that $N^{\alpha,\beta}$ is the number of samples that the n-th variable takes α or β. We next derive the number of total samples needed in order to have $N^{\alpha,\beta}$ samples for a given (α,β) pair. Since $\mathcal{D}(\mathcal{W},\Theta)$ is δ-unbiased with $\delta = e^{-2\lambda(\mathcal{W},\Theta)}/k$, for $Z \sim \mathcal{D}(\mathcal{W},\Theta)$, we have $\mathbb{P}[Z_n \in \{\alpha,\beta\}|Z_{-n}] \geq 2\delta$. By the Chernoff bound, if the total number of samples satisfies $N = O(N^{\alpha,\beta}/\delta + \log(1/\rho'')/\delta)$, then with probability at least $1-\rho''$, we have $N^{\alpha,\beta}$ samples for a given (α,β) pair.

To ensure that (51) holds for all (α,β) pairs with probability at least $1-\rho/n$, we can set $\rho' = \rho/(nk^2)$ and $\rho'' = \rho/(nk^2)$ and take a union bound over all (α,β) pairs. The total number of samples required is $N = O(\lambda^2 k^4 \exp(14\lambda)\ln(nk/\rho)/\epsilon^4)$.

We have shown that (47) holds for the n-th variable with probability at least $1-\rho/n$. By the discussion at the beginning of the proof, Theorem 2 then follows by a union bound over the n variables. □

H Mirror descent algorithms for constrained logistic regression

Algorithm 3 gives a mirror descent algorithm for the following ℓ_1-constrained logistic regression:

$$\min_{w\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N} \ln(1+e^{-y^i\langle w,x^i\rangle}) \qquad \text{s.t. } \|w\|_1 \leq W_1. \tag{52}$$

We use the doubling trick to expand the dimension and re-scale the samples (Step 1-4). Now the original problem becomes a logistic regression problem over a probability simplex: $\Delta_{2n+1} = \{w \in \mathbb{R}^{2n+1} : \sum_{i=1}^{2n+1} w_i = 1, w_i \geq 0, \forall i \in [2n+1]\}$.

$$\min_{w\in\Delta_{2n+1}} \frac{1}{N}\sum_{i=1}^{N} -\hat{y}^i \ln(\sigma(\left\langle w, \hat{x}^i\right\rangle)) - (1-\hat{y}^i)\ln(1-\sigma(\left\langle w,\hat{x}^i\right\rangle)), \tag{53}$$

where $(\hat{x}^i, \hat{y}^i) \in \mathbb{R}^{2n+1} \times \{0,1\}$. We then (Step 4-11) follow the standard simplex setup for mirror descent algorithm (see Section 5.3.3.2 of [BTN13]). Specifically, the negative entropy is used as the distance generating function (aka the mirror map). The projection step (Step 9) can be done by a simple ℓ_1 normalization operation. After that, we transform the solution back to the original space (Step 12).

Algorithm 3 Mirror descent algorithm for ℓ_1-constrained logistic regression

Input: $\{(x^i, y^i)\}_{i=1}^N$ where $x^i \in \{-1,1\}^n$, $y^i \in \{-1,1\}$; constraint on the ℓ_1 norm $W_1 \in \mathbb{R}_+$; number of iterations T.
Output: $\bar{w} \in \mathbb{R}^n$.

```
for each sample $i \in [N]$ do
    $\hat{x}^i \leftarrow [x^i, -x^i, 0] \cdot W_1, \quad \hat{y}^i \leftarrow (y^i+1)/2$        ▷ Form samples $(\hat{x}^i, \hat{y}^i) \in \mathbb{R}^{2n+1} \times \{0,1\}$.
$w^1 \leftarrow [\frac{1}{2n+1}, \frac{1}{2n+1}, \cdots, \frac{1}{2n+1}] \in \mathbb{R}^{2n+1}$        ▷ Initialize $w$ as the uniform distribution.
$\gamma \leftarrow \frac{1}{2W_1}\sqrt{\frac{2\ln(2n+1)}{T}}$        ▷ Set the step size.
for each iteration $t \in [T]$ do
    $g^t \leftarrow \frac{1}{N}\sum_{i=1}^N (\sigma(\langle w^t, \hat{x}^i \rangle) - \hat{y}^i)\hat{x}^i$        ▷ Compute the gradient.
    $w_i^{t+1} \leftarrow w_i^t \exp(-\gamma g_i^t)$, for $i \in [2n+1]$        ▷ Coordinate-wise update.
    $w^{t+1} \leftarrow w^{t+1}/\|w^{t+1}\|_1$        ▷ Projection step.
$\bar{w} \leftarrow \sum_{t=1}^T w^t/T$        ▷ Aggregate the updates.
$\bar{w} \leftarrow (\bar{w}_{1:n} - \bar{w}_{(n+1):2n}) \cdot W_1$        ▷ Transform $\bar{w}$ back to $\mathbb{R}^n$ and the actual scale.
```

Algorithm 4 gives a mirror descent algorithm for the following $\ell_{2,1}$-constrained logistic regression:

$$\min_{w \in \mathbb{R}^{n \times k}} \frac{1}{N}\sum_{i=1}^N \ln(1 + e^{-y^i\langle w, x^i \rangle}) \qquad \text{s.t. } \|w\|_{2,1} \le W_{2,1}. \tag{54}$$

For simplicity, we assume that $n \ge 3$[6]. We then follow Section 5.3.3.3 of [BTN13] to use the following function as the mirror map $\Phi : \mathbb{R}^{n \times k} \to \mathbb{R}$:

$$\Phi(w) = \frac{e\ln(n)}{p}\|w\|_{2,p}^p, \quad p = 1 + 1/\ln(n). \tag{55}$$

The update step (Step 8) can be computed efficiently in $O(nk)$ time, see the discussion in Section 5.3.3.3 of [BTN13] for more details.

I Proof of Theorem 3 and Theorem 4

The proof relies on the following convergence result of the mirror descent algorithms given in Appendix H.

Lemma 11. *Let $\hat{\mathcal{L}}(w) = \frac{1}{N}\sum_{i=1}^N \ln(1+e^{-y^i\langle w, x^i\rangle})$ be the empirical loss. Let $\hat{w}$ be a minimizer of the ERM defined in (52). The output $\bar{w}$ of Algorithm 3 satisfies*

$$\hat{\mathcal{L}}(\bar{w}) - \hat{\mathcal{L}}(\hat{w}) \le 2W_1\sqrt{\frac{2\ln(2n+1)}{T}}. \tag{56}$$

Similarly, let $\hat{w}$ be a minimizer of the ERM defined in (54). Then the output $\bar{w}$ of Algorithm 4 satisfies

$$\hat{\mathcal{L}}(\bar{w}) - \hat{\mathcal{L}}(\hat{w}) \le O(1) \cdot W_{2,1}\sqrt{\frac{\ln(n)}{T}}. \tag{57}$$

[6] For $n \le 2$, we need to switch to a different mirror map, see Section 5.3.3.3 of [BTN13] for more details.

Algorithm 4 Mirror descent algorithm for $\ell_{2,1}$-constrained logistic regression

Input: $\{(x^i, y^i)\}_{i=1}^N$ where $x^i \in \{0,1\}^{n\times k}$, $y^i \in \{-1,1\}$; constraint on the $\ell_{2,1}$ norm $W_{2,1} \in \mathbb{R}_+$; number of iterations T.
Output: $\bar{w} \in \mathbb{R}^{n\times k}$.

```
1: for each sample $i \in [N]$ do
2:     $\hat{x}^i \leftarrow x^i \cdot W_{2,1}$,   $\hat{y}^i \leftarrow (y^i + 1)/2$      ▷ Form samples $(\hat{x}^i, \hat{y}^i) \in \mathbb{R}^{n\times k} \times \{0,1\}$.
3: $w^1 \leftarrow [\frac{1}{n\sqrt{k}}, \frac{1}{n\sqrt{k}}, \cdots, \frac{1}{n\sqrt{k}}] \in \mathbb{R}^{n\times k}$      ▷ Initialize $w$ as a constant matrix.
4: $\gamma \leftarrow \frac{1}{2W_{2,1}}\sqrt{\frac{e\ln(n)}{T}}$      ▷ Set the step size.
5: for each iteration $t \in [T]$ do
6:     $g^t \leftarrow \frac{1}{N}\sum_{i=1}^N(\sigma(\langle w^t, \hat{x}^i\rangle) - \hat{y}^i)\hat{x}^i$      ▷ Compute the gradient.
7:     $w^{t+1} \leftarrow \arg\min_{\|w\|_{2,1}\le 1} \Phi(w) - \langle \nabla\Phi(w^t) - \gamma g^t, w\rangle$      ▷ $\Phi(w)$ is defined in (55).
8: $\bar{w} \leftarrow (\sum_{t=1}^T w^t/T) \cdot W_{21}$      ▷ Aggregate the updates.
```

Lemma 11 follows from the standard convergence result for mirror descent algorithm (see, e.g., Theorem 4.2 of [Bub15]), and the fact that the gradient g^t in Step 6 of Algorithm 3 satisfies $\|g^t\|_\infty \le 2W_1$ (reps. the gradient g^t in Step 6 of Algorithm 4 satisfies $\|g^t\|_\infty \le 2W_{2,1}$). This implies that the objective function after rescaling the samples is $2W_1$-Lipschitz w.r.t. $\|\cdot\|_1$ (reps. $2W_{2,1}$-Lipschitz w.r.t. $\|\cdot\|_{2,1}$).

We are ready to prove Theorem 3, which is restated below.

Theorem. *In the setup of Theorem 1, suppose that the ℓ_1-constrained logistic regression in Algorithm 1 is optimized using the mirror descent algorithm given in Appendix H. Given $\rho \in (0,1)$ and $\epsilon > 0$, if the number of mirror descent iterations satisfies $T = O(\lambda^2 \exp(12\lambda)\ln(n)/\epsilon^4$, and the number of i.i.d. samples satisfies $N = O(\lambda^2\exp(12\lambda)\ln(n/\rho)/\epsilon^4)$, then (6) still holds with probability at least $1-\rho$. The total run-time of Algorithm 1 is $O(TNn^2)$.*

Proof. We first note that in the proof of Theorem 1, we only use $\hat{w}$ in order to apply the result from Lemma 1. In the proof of Lemma 1, there is only one place where we use the definition of $\hat{w}$: the inequality (b) in (29). As a result, if we can show that (29) still holds after replacing $\hat{w}$ by $\bar{w}$, i.e.,

$$\mathcal{L}(\bar{w}) \le \mathcal{L}(w^*) + O(\gamma), \tag{58}$$

then Lemma 1 would still hold, and so is Theorem 1.

By Lemma 11, if the number of iterations $T = O(W_1^2\ln(n)/\gamma^2)$, then

$$\hat{\mathcal{L}}(\bar{w}) - \hat{\mathcal{L}}(\hat{w}) \le \gamma. \tag{59}$$

As a result, we have

$$\mathcal{L}(\bar{w}) \overset{(a)}{\le} \hat{\mathcal{L}}(\bar{w}) + \gamma \overset{(b)}{\le} \hat{\mathcal{L}}(\hat{w}) + 2\gamma \overset{(c)}{\le} \hat{\mathcal{L}}(w^*) + 2\gamma \overset{(d)}{\le} \mathcal{L}(w^*) + 3\gamma, \tag{60}$$

where (a) follows from (27), (b) follows from (59), (c) follows from the fact that $\hat{w}$ is the minimizer of $\hat{\mathcal{L}}(w)$, and (d) follows from (28). The number of mirror descent iterations needed for (58) to hold is $T = O(W_1^2\ln(n)/\gamma^2)$. In the proof of Theorem 1, we need to set $\gamma = O(1)\epsilon^2\exp(-6\lambda)$ (see the proof following (44)), so the number of mirror descent iterations needed is $T = O(\lambda^2\exp(12\lambda)\ln(n)/\epsilon^4)$.

To analyze the runtime of Algorithm 1, note that for each variable in $[n]$, transforming the samples takes $O(N)$ time, solving the ℓ_1-constrained logisitic regression via Algorithm 3 takes $O(TNn)$ time, and updating the edge weight estimate takes $O(n)$ time. Forming the graph $\hat{G}$ over n nodes takes $O(n^2)$ time. The total runtime is $O(TNn^2)$. □

The proof of Theorem 4 is identical to that of Theorem 3 and is omitted here. The key step is to show that (58) holds after replacing $\hat{w}$ by $\bar{w}$. This can be done by using the convergence result in Lemma 11 and applying the same logic in (60). The runtime of Algorithm 2 can be analyzed in the same way as above. The $\ell_{2,1}$-constrained logistic regression dominates the total runtime. It requires $O(TN^{\alpha,\beta}nk)$ time for each pair (α, β) and each variable in $[n]$, where $N^{\alpha,\beta}$ is the subset of samples that a given variable takes either α or β. Since $N \geq kN^{\alpha,\beta}$, the total runtime is $O(TNn^2k^2)$.