Proceedings of Machine Learning Research vol 99:1–21, 2019 32nd Annual Conference on Learning Theory

Learning Ising Models with Independent Failures

Surbhi Goel SURBHI@CS.UTEXAS.EDU
University of Texas at Austin

Daniel M. Kane DAKANE@UCSD.EDU
University of California San Diego

Adam R. Klivans KLIVANS@CS.UTEXAS.EDU
University of Texas at Austin

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We give the first efficient algorithm for learning the structure of an Ising model that tolerates independent failures; that is, each entry of the observed sample is missing with some unknown probability p. Our algorithm matches the essentially optimal runtime and sample complexity bounds of recent work for learning Ising models due to Klivans and Meka (2017).

We devise a novel unbiased estimator for the gradient of the *Interaction Screening Objective* (ISO) due to Vuffray et al. (2016) and apply a stochastic multiplicative gradient descent algorithm to minimize this objective. Solutions to this minimization recover the neighborhood information of the underlying Ising model on a node by node basis.

Keywords: graphical models, Ising models, structure learning, efficient algorithms, missing data, unbiased estimators

1. Introduction

Ising models are fundamental undirected binary graphical models that capture pair-wise dependencies between the input variables. They are well-studied in the literature and have applications in a large number of areas such as physics, computer vision and statistics (Jaimovich et al. (2006); Koller et al. (2009); Choi et al. (2010); Marbach et al. (2012)). One of the core problems in understanding graphical models is *structure learning*, that is, recovering the structure of the underlying graph given access to random samples from the distribution. Developing efficient algorithms for structure learning is a heavily-studied topic, especially for the case when the underlying graph is sparse or has bounded degree (Dasgupta (1999); Lee et al. (2007); Bresler et al. (2008); Ravikumar et al. (2010); Bresler et al. (2014); Bresler (2015); Vuffray et al. (2016); Hamilton et al. (2017); Klivans and Meka (2017); Wu et al. (2018)). Klivans and Meka (2017) were the first to give an algorithm for learning the structure of Ising models (and more generally higher order MRFs) with essentially optimal runtime and sample complexity.

The focus of this paper is the setting where the samples drawn from the Ising model are corrupted by noise. More specifically, we consider the *independent failure model* where each entry of each sample is independently corrupted – missing or flipped — with some constant possibly unknown probability p. Such a scenario can be expected, for example, in a sensor network, where sensors fail to report data occasionally due to internal failures. In addition to being a natural question, structure learning in this model has been specifically posed as an open problem by Chen.

1.1. Our Result

The main contribution of our paper is a simple stochastic algorithm for learning the structure of an Ising model in the presence of corruptions with near optimal sample and time complexity:

Theorem 1 (Informal version) *Consider an Ising model with dependency graph G over n vertices with minimum edge weight β. Further assume that the sum of absolute values of the weights of outgoing edges from each vertex is bounded by λ. Given corrupted draws from the Ising model such that each entry is missing or flipped with known probability $p \in (0, 1)$, there exists an algorithm that recovers the underlying structure in time $\widetilde{O}(n^2)$ using at most $\lambda^4 \beta^{-4} \log(n) \exp(C\lambda)$ samples where C[1] depends on p.*

For missing data, we extend our approach to the setting where p is unknown. We show that using only $O(\log n)$ fresh samples to estimate these probabilities suffices for the guarantees to hold. Our results can also be easily extended to other similar noise models such as independent block failures where instead of each entry being flipped, fixed subsets of entries are simultaneously missing or flipped. Our work is the first efficient algorithm for learning the structure of Ising models in the presence of independent failures; both our running time and sample complexity are essentially optimal (matching recent work due to Klivans and Meka (2017)).

1.2. Our Approach

Our approach has three main components that we briefly explain here.

Optimization Problem: We follow the "nodewise-regression" approach of recovering the graph by solving an optimization problem to identify the neighborhood of each vertex. To do so, we minimize the Interaction Screening Objective (ISO) proposed by Vuffray et al. (2016) over the l_1-norm constrained ball. The ISO satisfies a property known as *restricted strong convexity* which allows us to recover the underlying edge weights from an approximate minimum solution.

Minimization Procedure: Instead of using convex programming as in Vuffray et al. (2016), we minimize ISO using *stochastic multiplicative gradient descent* (SMG) due to Kakade et al. (2008). SMG is a stochastic, multiplicative-weight update algorithm and therefore runs on the simplex (modeling a probability distribution). As such, we need to transform our optimization problem to the simplex. The main motivation for using SMG is two-fold: 1) it can handle l_1-constraints, 2) it requires access to only an unbiased, bounded estimator of the gradient to give convergence guarantees.

Unbiased Gradient Construction: Despite independent failures in our sample, we are able to construct an unbiased estimator of the gradient of the (modified) ISO on the simplex. Our constructor is able to exploit the *decomposability* of the ISO, as it is an exponential of a linear function. This allows us to use the independence property of the failures to come up with a simple unbiased estimator. The techniques used for this construction may be of independent interest.

1. In the missing-data setting, C scales as $\frac{1}{1-p}$. For $p < 1/2$, C is at most 10. We refer the reader to the proof of Theorem 10 for the exact dependence on p.

1.3. Related Work

Ising Models. Structure learning for Markov Random Fields has been studied since the 1960s. For example, Chow and Liu (1968) gave a greedy algorithm for undirected graphical models assuming the underlying graph is a tree. There have been many works for learning Ising models under various assumptions on the structure of the underlying graph (e.g., Lee et al. (2007); Yuan and Lin (2007); Ravikumar et al. (2010); Yang et al. (2012)). For Ising models specifically, the first assumption-free result (that is, no assumptions are made on the underlying graph other than sparsity) was given by Bresler (2015). One drawback of Bresler's work, however, is that the sample complexity has a doubly exponential dependence on the sparsity of the underlying graph. This was improved by Vuffray et al. (2016) where they used general-purpose tools for convex programming to minimize a certain objective function, but the running time of their approach was suboptimal. Subsequently, Klivans and Meka (2017) gave a multiplicative weight update algorithm called *Sparsitron* that achieved near-optimal sample complexity (essentially matching known information-theoretic lower bounds) and near-optimal running time (under a computational hardness assumption for the *light-bulb problem* Valiant (2015)). Recently, Wu et al. (2018) gave a different proof of Klivans and Meka (2017) using l_1-regularized logistic regression.

Learning Ising Models with Noise. None of the works mentioned above handle the setting where there is noise in the observed data. The problem of structure learning in the independent failure model (specifically missing data) was raised by Chen. The only work we are aware of that obtains positive results in this model is due to Hamilton et al. (2017), who applied a broad generalization of the work of Bresler (2015). Their sample complexity, however, has a doubly exponential dependence on the sparsity of the underlying graph. Our result gives a singly exponential dependence on the sparsity (known to be information-theoretically optimal).

Robust Learning of Ising Models. Lindgren et al. (2018) studied structure learning in a different noise model motivated by recent results in robust learning. In their work, an adversary is allowed to arbitrarily corrupt some η fraction of a training set drawn from the underlying distribution. They showed that if an adversary is allowed to corrupt even an exponentially small fraction of the samples, then no algorithm is robust. They further showed that their bounds are tight by proving robustness of *Sparsitron* against exponentially small adversarial corruption. The adversarial model is incomparable to the model studied here: in the independent failure model, *every* example will (on average) have a p fraction of the entries missing (or flipped).

Learning with Missing Data. For general learning problems, various methods have been proposed to handle missing data including heuristics and maximum likelihood methods. In the context of high-dimensional sparse *linear* regression, Loh and Wainwright (2011) proposed simple estimators to handle missing data based on solving optimization problems, and further showed that simple gradient descent recovers close to optimum solution despite non-convexity. For distribution learning, Shah and Song (2018) recently gave the first positive results for learning mixtures of gaussians with missing data with optimal sample/runtime complexity. It is not immediately clear how to use either of these techniques for our problem.

1.4. Notation

For vector $x \in \mathbb{R}^n$, x_{-i} denotes $(x_j : j \neq i)$ and x_S denotes $(x_i : i \in S)$. $||\cdot||_p$ denoted the l_p-norm. We denote the different distance/divergence metrics for distributions as follows: $\mathsf{KL}(P||Q)$

denotes the Kullback-Lieber divergence between probability distributions P and Q, and $d_{\mathsf{TV}}(P,Q)$ denotes the total variation distance between P and Q. We denote the l_1 ball of radius R using $\mathsf{B}(R,n) := \{x \in \mathbb{R}^n : ||x||_1 \leq R\}$ and the simplex with radius R using $\Delta(R,n) := \{x \in \mathbb{R}^n : x \geq 0, \sum_{i=1}^n x_i = R\}$.

2. Preliminaries

Definition 2 (Ising model) *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with $A_{ii} = 0$ for all i and $\theta \in \mathbb{R}^n$ be the mean-field vector. Let $G = (V,E)$ be an undirected graph with $V = [n]$ such that $(i,j) \in E$ if and only if $A_{ij} \neq 0$. The n-variable Ising model with underlying dependency graph G is a distribution $\mathcal{D}(A,\theta)$ defined on $\{-1,1\}^n$ such that*

$$\Pr[Z = z] = \frac{1}{\mathcal{Z}} \exp\left(\sum_{(i,j)\in E} A_{ij} z_i z_j + \sum_{i\in V} \theta_i z_i\right)$$

where $\mathcal{Z} = \sum_{z\in\{-1,1\}^n} \exp\left(\sum_{(i,j)\in E} A_{ij} z_i z_j + \sum_{i\in V} \theta_i z_i\right)$ is the normalizing factor. We denote the minimum edge weight $\beta(A) := \min_{(i,j)\in E} |A_{ij}|$ and the width of the model by $\lambda(A,\theta) := \max_{i\in V}\left(\sum_{j|(i,j)\in E} |A_{ij}| + |\theta_i|\right)$.

We will assume that the minimum edge weight is at least $\beta \leq \beta(A,\theta)$ and the model is width bounded by $\lambda \geq \lambda(A,\theta)$. For ease of presentation, we will suppress notation and denote $\mathcal{D}(A,\theta)$ by $\mathcal{D}$. A useful property of bounded width Ising models is that the conditional distributions are bounded away from 0 and 1. More formally,

Lemma 3 (Bresler (2015)) *For any node $v \in V$, subset $S \subseteq V\backslash\{v\}$, and any fixed configuration z_S of the subset S,*

$$\min\{\Pr[Z_v = 1 | Z_S = z_S], \Pr[Z_v = -1 | Z_S = z_S]\} \geq \frac{1}{2}\exp(-2\lambda).$$

In this paper, we are interested in *structure learning*, that is, recovering the edges of the dependency graph of an unknown Ising model given independent draws from it. We will focus on recovery in the presence of corruptions in the observed samples. The model of corruptions we study in our paper is known as the *independent failures models*: corruptions of each variable are independent of all the other variables. More specifically, we will consider the following two special cases of independent failures:

1. *Missing Data*: In this setting, for sample z we will instead observe x such that for all i, with probability p_i, $x_i =?$ (missing) otherwise $x_i = z_i$.

2. *Flipped Data*: In this setting, for sample z we will instead observe x such that for all i, with probability p_i, $x_i = -z_i$ (flipped) otherwise $x_i = z_i$.

3. Main Approach

To recover the underlying graph, we will show how to reconstruct the neighborhood of each vertex by solving a convex constrained minimization problem. We will first describe the optimization problem and then show how to minimize the same using a stochastic first-order method assuming access to an unbiased estimator of the gradient. Subsequently we will show how to recover the neighborhood from the optimization solution. Lastly, we will detail the construction of the unbiased estimators for the two given noise models.

Note: For ease of presentation, we will consider Ising models with zero mean-field ($\theta = 0$). For details on the non-zero mean-field case, we refer the reader to Appendix B.

3.1. Optimization Problem

Consider the neighborhood of a fixed vertex (WLOG we choose n, we can do the same analysis for any vertex). Vuffray et al. (2016) proposed to minimize the Interaction Screening Objective (ISO) as follows:

Optimization Problem 1:

$$\begin{aligned} \min_{v \in \mathbb{R}^{n-1}} \quad & S(v) := \mathbb{E}_{Z \sim \mathcal{D}} \left[\exp\left(- \sum_{j=1}^{n-1} v_j Z_n Z_j \right) \right] \\ \text{subject to} \quad & ||v||_1 \leq \lambda. \end{aligned}$$

Vuffray et al. (2016) studied the empirical version of Optimization Problem 1: the objective is computed using m samples $\{z^1, \ldots, z^m\}$ drawn from the Ising model as $\hat{S}(v) := \frac{1}{m} \sum_{i=1}^{m} \exp\left(- \sum_{j=1}^{n-1} v_j z_n^i z_j^i \right)$. They proved various useful properties of $\hat{S}$ that directly extend to S. Firstly, they showed that the optimal solution to Optimization Problem 1 captures the edge weights of the neighborhood of n in G. More formally,

Lemma 4 (Vuffray et al. (2016)) *Let $v^* \in \mathbb{R}^{n-1}$ be such that for $i \in [n-1]$, $v_i^* = A_{ni}$ if $(i, n) \in E$ else 0. Then we have, $\nabla S(v^*) = 0$ and v^* is a global minimum of Optimization Problem 1.*

Further they proved that S satisfies a property known as *restricted strong convexity* (RSC) which enables us to recover a vector close to v^* by approximately minimizing the objective. Here we give a stronger version of their result which holds more generally.

Lemma 5 *[RSC for S] For all $v \in \mathsf{B}(\lambda, n-1)$,*

$$S(v) - S(v^*) \geq \nabla S(v^*)^T (v - v^*) + \frac{\exp(-3\lambda)}{1 + \lambda} ||v - v^*||_\infty^2$$

Note: Our proof technique improves on their result, as we work directly with the l_∞ norm, avoiding the use of the l_2 norm in their proof (c.f. Lemma 7 and 8 Vuffray et al. (2016)). Using our analysis in their proof improves their sample complexity bounds for structure learning from $\max(d, \beta^{-2}) d^3 e^{O(\beta d)} \log n$ to $\max(d^2, \beta^{-2}) d^2 e^{O(\beta d)} \log n$ where d is the maximum degree of each node in G. This improvement is highlighted when $\beta d < 1$ (where the improvement is by a factor of d).

With the above lemma in hand, we can show that small loss implies closeness to v^* and hence recovery of the edge weights of the neighborhood of n. More formally, if $S(v) - S(v^*) \leq \epsilon$ then $||v - v^*||_\infty^2 = \max_{i \in [n-1]} |v_i - A_{ni}|^2 \leq (1 + \lambda) \exp(3\lambda)\epsilon$.

3.2. Minimizing ISO with l_1 Constraint

To solve Optimization Problem 1, we will use an algorithm due to Kakade et al. (2008) called Stochastic Multiplicative Gradient Descent (SMG) (see Algorithm 3.2). SMG is a multiplicative-weight update algorithm that optimizes over the simplex instead of the l_1-constraint ball. As such, we will need to reduce our problem from the l_1 ball to the simplex to obtain the appropriate guarantees.

Algorithm 1 Stochastic Multiplicative Gradient Descent

Input: l_1-bound W, dimension k and convex function c over $\Delta(W, k)$.
Set $u_i^1 := W/k$ for all i.
for $t = 1$ **to** T **do**
  For unbiased estimate of gradient g^t of $c(u^t)$ and $C^t := \frac{(u^t)^T g^t}{W}$
  For all i, update $u_i^{t+1} := u_i^t \left(1 - \eta g_i^t + \eta C^t\right)$
end for
Output: $\bar{u} = \frac{1}{T} \sum_{i=1}^{T} u^t$

Theorem 6 (Kakade et al. (2008)) *Let u^* be an optimal solution of $\min_{u \in \Delta(W,k)} c(u)$ for convex function c on $\Delta(W, k)$. For SMG run on c, as long as $\mathbb{E}[g^t | \mathcal{H}^{t-1}] = \nabla c(u^t)$ where $\mathcal{H}^{t-1}$ denotes history of previous iterations and $||\nabla c(u^t)||_\infty, ||g^t||_\infty \leq B$ for all u on the simplex and for all iterations $t \leq T$, then with probability $1 - \delta$ for $\eta = \frac{1}{2B}\sqrt{\frac{\log k}{T}}$ and $T \geq 16 \log k$,*

$$c(\bar{u}) \leq c(u^*) + 4BW \left(\sqrt{\frac{\log k}{T}} + \sqrt{\frac{2}{T} \log \frac{1}{\delta}} \right).$$

Note: Since the SMG analysis only appears in a set of lecture notes (and there are some minor errors in the writeup), for completeness we include the proof of the theorem in the appendix.

To apply SMG, we convert our optimization problem's constraint from the l_1 ball to the simplex. To do so, we define the following mappings:

$$\Pi_{B \to \Delta}^k : \mathsf{B}(W, k) \to \Delta(W, 2k{+}1), \quad \Pi_{B \to \Delta}^k(w)_i = \begin{cases} W - ||w||_1 & \text{if } i = 2k + 1 \\ w_i & \text{if} i \in \{1, \ldots, k\}, w_i \geq 0 \\ -w_{i-k} & \text{if } i \in \{k + 1, \ldots, 2k\}, w_{i-k} < 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$\Pi_{\Delta \to B}^k : \Delta(W, 2k + 1) \to \mathsf{B}(W, k), \quad \Pi_{\Delta \to B}^k(u)_i = u_i - u_{i+k} \text{ for } i \in \{1, \ldots, k\}$$

It is not hard to verify that these are valid mappings from ball to simplex and vice-versa. Using the given transformation, the optimization problem on the simplex is as follows:

Optimization Problem 2:

$$\min_{w\in\mathbb{R}^{2n-1}} \quad \widetilde{S}(w) := S(\Pi_{\Delta\to B}^{n-1}(w)) = \mathbb{E}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j=1}^{n-1}(w_j - w_{n-1+j})Z_nZ_j\right)\right]$$

$$\text{subject to} \quad ||w||_1 = \lambda, w \geq 0.$$

Note that the above loss is also convex and similar to Lemma 4, we can show the following,

Lemma 7 *Let $w^* = \Pi_{B\to\Delta}^{n-1}(v^*)$, then $\nabla\widetilde{S}(w^*) = 0$ and w^* is a global minimum of Optimization Problem 2.*

Using SMG we can thus solve Optimization Problem 2 as long as we have access to an unbiased estimator of the gradient.

3.3. Unbiased Estimator of Gradient

Let v be a candidate solution of Optimization Problem 2. Since we have missing/flipped data, it is not clear how to compute the loss or the gradient at the point v, and the gradient is required to execute a step of the SMG algorithm. We will show, however, that it is possible to construct an unbiased estimator of the gradient with bounded l_∞ norm for known p_is. We then show that our estimates are sufficiently strong to plug-in to the SMG analysis.

3.3.1. MISSING DATA

In the missing-data model, instead of getting samples from $\mathcal{D}$, we instead get samples from a corrupted distribution $\mathcal{D}_{\mathsf{miss}}$ as follows:

- Let $Z \sim \mathcal{D}$. Let $C_1 \sim \mathsf{Ber}(1-p_1), \ldots, C_n \sim \mathsf{Ber}(1-p_n)$.
- Output X such that for all i, $X_i = C_iZ_i$ (replacing ? by 0).

Here $\mathsf{Ber}(p)$ denotes the distribution of a Bernoulli random variable with probability p.

The main intuition of our estimator can be made clear with the following example. Suppose we wish to construct an unbiased estimator for function $f(Z_i)$ then consider $g(X_i) = \frac{f(X_i)-pf(0)}{1-p}$. Then it is not hard to see that,

$$\begin{aligned}\mathbb{E}_{X_i}\left[\frac{f(X_i)-pf(0)}{1-p}\middle|Z_i\right] &= \mathop{\mathbb{E}}_{C_i\sim\mathsf{Ber}(1-p_i)}\left[\frac{f(C_iZ_i)-pf(0)}{1-p}\middle|Z_i\right]\\ &= (1-p)\times\frac{f(Z_i)-pf(0)}{1-p} + p\times f(0) = f(Z_i).\end{aligned}$$

Since ISO is a product function in $Z_1, \ldots, Z_{n-1}$, we can apply the above estimator to each term of the product. For Z_n we use the above idea on the so formed product of estimators. This allows us to construct the following estimators, $\forall i \in [n-1]$,

$$g^i_{\mathsf{miss}}(w;X) = -\frac{X_n}{1-p_n}\times\frac{\exp\left(-(w_i - w_{n-1+i})X_nX_i\right)X_i}{1-p_i}\times\prod_{j\neq i,n}\frac{\exp\left(-(w_j - w_{n-1+j})X_nX_j\right)-p_j}{1-p_j}.$$

The following lemma shows how to use g_{miss} to construct an unbiased estimator of the gradient of $\widetilde{S}$.

Lemma 8 *Consider estimator* $G_{\mathsf{miss}}(w;X) = \sum_{i=1}^{n-1} g^i_{\mathsf{miss}}(w;X)(e^i - e^{n-1+i})$ *where* $e^i \in \mathbb{R}^{2n-1}$ *is the indicator vector for coordinate* i*. Then for all fixed* w*,*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{miss}}[G_{\mathsf{miss}}(w;X)] = \nabla\widetilde{S}(w).$$

Also, for all $X \in \{0,1\}^n$ *and* $w \in \Delta B(\lambda, 2n-1)$*,*

$$||G_{\mathsf{miss}}(w;X)||_\infty \leq \frac{1}{(1-p_{\max})^2}\exp\left(\frac{\lambda}{1-p_{\max}}\right)$$

where $p_{\max} = \max_i p_i$*.*

Proof Observe that the gradient of $\widetilde{S}(w)$ with respect to w can be computed as follows. We have $\nabla\widetilde{S}(w)_{2n-1} = 0$ since $\widetilde{S}$ does not depend on w_{2n-1}, and

$$\forall i \in [n-1], \nabla\widetilde{S}(w)_i = -\nabla\widetilde{S}(w)_{i+n-1} = -\mathbb{E}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j=1}^{n-1}(w_j - w_{n-1+j})Z_nZ_j\right)Z_nZ_i\right].$$

For ease of presentation, let $v = \Pi^{n-1}_{\Delta\to B}(w)$ that is, $v_i = w_i - w_{n-1+i}$ for $i \in [n-1]$. Taking expectation of g_{miss} over $\mathcal{D}_{\mathsf{miss}}$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{miss}}}[g^i_{\mathsf{miss}}(w;X)]\\
&= -\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{miss}}}\left[\frac{X_n}{1-p_n}\times\frac{\exp\left(-v_iX_nX_i\right)X_i}{1-p_i}\times\prod_{j\neq i,n}\frac{\exp\left(-v_jX_nX_j\right)-p_i}{1-p_i}\right]\\
&= -\mathop{\mathbb{E}}_{Z\sim\mathcal{D},C_n\sim\mathsf{Ber}(1-p_n)}\left[\frac{C_nZ_n}{1-p_n}\times\mathop{\mathbb{E}}_{C_i\sim\mathsf{Ber}(1-p_i)}\left[\frac{\exp\left(-v_iC_nZ_nC_iZ_i\right)C_iZ_i}{1-p_i}\right]\times\right.\\
&\qquad\left.\prod_{j\neq i,n}\mathop{\mathbb{E}}_{C_j\sim\mathsf{Ber}(1-p_j)}\left[\frac{\exp\left(-v_jC_nZ_nC_jZ_j\right)-p_j}{1-p_j}\right]\right]\\
&= -\mathop{\mathbb{E}}_{\substack{Z\sim\mathcal{D}\\C_n\sim\mathsf{Ber}(1-p_n)}}\left[\frac{C_nX_n}{1-p_n}\times\exp\left(-v_iC_nZ_nZ_i\right)Z_i\times\prod_{j\neq i,n}\exp\left(-v_jC_nZ_nZ_j\right)\right]\\
&= -\mathop{\mathbb{E}}_{\substack{Z\sim\mathcal{D}\\C_n\sim\mathsf{Ber}(1-p_n)}}\left[\frac{\exp\left(-\sum_{j\neq n}v_jC_nZ_nZ_j\right)C_nZ_nZ_i}{1-p_n}\right]\\
&= -\mathop{\mathbb{E}}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j\neq n}v_jZ_nZ_j\right)Z_nZ_i\right] = \nabla\widetilde{S}(w)_i.
\end{aligned}$$

The above follows from observing the following facts: 1) $\mathbb{E}_{C\sim\mathsf{Ber}(1-p)}\left[\frac{\exp(CX)-p}{1-p}\right] = \exp(X)$ as long as X is independent of C, 2) $\mathbb{E}_{C\sim\mathsf{Ber}(1-p)}\left[\frac{\exp(CX)CY}{1-p}\right] = \exp(X)\,Y$ as long as X, Y are independent of C.

The last thing we need is that the above estimator has bounded l_∞ norm for all $w \in \Delta(\lambda, 2n-1)$ ($v \in B(\lambda, n-1)$). Observe that

$$\left|\frac{\exp(a)-p}{1-p}\right| = \frac{\left|1-p+\sum_{i=1}^{\infty}\frac{a^i}{i!}\right|}{1-p} \leq 1 + \sum_{i=1}^{\infty}\frac{|a|^i}{(1-p)i!} \leq 1 + \sum_{i=1}^{\infty}\frac{\left(\frac{a}{1-p}\right)^i}{i!} = \exp\left(\frac{|a|}{1-p}\right).$$

Using the above property,

$$\begin{aligned}
|g^i_{\mathsf{miss}}(w;X)| &\leq \left|\frac{X_n}{1-p_n}\right| \times \left|\frac{\exp\left(-v_iX_nX_i\right)X_i}{1-p_i}\right| \times \prod_{j\neq i,n}\left|\frac{\exp\left(-v_jX_nX_j\right)-p_i}{1-p_i}\right| \\
&\leq \frac{\exp\left(|v_i|\right)}{(1-p_{\max})(1-p_{\max})}\prod_{j\neq i,n}\exp\left(\frac{|v_i|}{1-p_{\max}}\right) \\
&\leq \frac{1}{(1-p_{\max})^2}\exp\left(\frac{\lambda}{1-p_{\max}}\right).
\end{aligned}$$

Here the inequality follows from observing that $\sum_{i=1}^{n-1}|v_i| = ||v||_1 \leq \lambda$. Thus $||G_{\mathsf{miss}}(w;X)||_\infty \leq \frac{1}{(1-p_{\max})^2}\exp\left(\frac{\lambda}{1-p_{\max}}\right)$ for all $w \in \Delta(\lambda, 2n-1)$. ■

3.3.2. RANDOM-FLIPPED DATA

In the random-flipped data model, instead of getting samples from $\mathcal{D}$, we instead get samples from a corrupted distribution $\mathcal{D}_{\mathsf{flip}}$ as follows:

- Let $Z \sim \mathcal{D}$. Let $C_1 \sim \mathsf{Rad}(1-p_1), \ldots, C_n \sim \mathsf{Rad}(1-p_n)$.
- Output X such that for all i, $X_i = C_iZ_i$.

Here $\mathsf{Rad}(p)$ denotes the distribution of Rademacher variables with probability p, that is, distribution over $\{-1, 1\}$ where probability of drawing 1 is p.

Similar to the missing data case, we motivate our estimator with the following example. Suppose we wish to construct an unbiased estimator for any function $f(Z_i)$ then consider $g(X_i) = \frac{(1-p)f(X_i)-pf(-X_i)}{1-2p}$. Then it is not hard to see that,

$$\begin{aligned}
&\mathbb{E}_{X_i}\left[\frac{(1-p)f(X_i)-pf(-X_i)}{1-2p}\middle| Z_i\right] \\
&= \mathop{\mathbb{E}}_{C_i\sim\mathsf{Rad}(1-p_i)}\left[\frac{(1-p)f(C_iZ_i)-pf(-C_iZ_i)}{1-2p}\middle| Z_i\right] \\
&= (1-p)\times\frac{(1-p)f(Z_i)-pf(-Z_i)}{1-2p} + p\times\frac{(1-p)f(-Z_i)-pf(Z_i)}{1-2p} = f(Z_i).
\end{aligned}$$

Based on the above example, define $\sigma(p,a,x) = \frac{(1-p)\exp(ax)-p\exp(-ax)}{1-2p}$. Consider the following estimators, for all $i \in [n-1]$:

$$g^i_{\mathsf{flip}}(w;X) = -\frac{(1-p_n)h^i(w;X)+p_nh^i(-w;X)}{1-2p_n}$$

Algorithm 2 SMG with Missing/Flipped Data on Ising Models

Set $w_i^1 := \lambda/(2n-1)$ for all i.
if Missing Data **then**
  $G(w;X) := G_{\mathsf{miss}}(w;X)$
end if
if Flipped Data **then**
  $G(w;X) := G_{\mathsf{flip}}(w;X)$
end if
for $t = 1$ **to** T **do**
  Draw missing/flipped data sample X^t from $\mathcal{D}_{\mathsf{miss}}$/$\mathcal{D}_{\mathsf{flip}}$ respectively.
  Let $C^t := \frac{(w^t)^T g(w^t;X^t)}{\lambda}$
  For all i, update $w_i^{t+1} := w_i^t \left(1 - \eta G(w^t;X^t)_i + \eta C^t\right)$
end for
Output $\bar{w} = \frac{1}{T}\sum_{i=1}^T w^t$

$$\text{where } h^i(w;X) = X_n X_i \times \prod_{j\neq n} \sigma\left(p_j, -(w_j - w_{n-1+j})X_n, X_j\right).$$

The following lemma shows that g_{flip} is indeed an unbiased estimator of $\nabla\widetilde{S}(w)_i$.

Lemma 9 *Using the above, we construct estimator $G_{\mathsf{flip}}(w;X) = \sum_{i=1}^{n-1} g^i_{\mathsf{flip}}(w;X)(e^i - e^{n-1+i})$ where $e^i \in \mathbb{R}^{2n-1}$ is the indicator vector for coordinate i. Then for all fixed w,*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{flip}}}[G_{\mathsf{flip}}(w;X)] = \nabla\widetilde{S}(w).$$

Also, for all $X \in \{-1,1\}^n$ and $w \in \Delta B(\lambda, 2n-1)$,

$$||G_{\mathsf{flip}}(w;X)||_\infty \leq \frac{1}{(1-2p_{\max})^2}\exp\left(\frac{\lambda}{1-2p_{\max}}\right)$$

where $p_{\max} = \max_i p_i$.

We defer the proof to the appendix as it follows roughly the same ideas as of the missing-data model.

4. Main Result

Combining the techniques presented in the previous sections, our main algorithm (Algorithm 4) gives us the following guarantees.

Theorem 10 *For known constants p_i, given samples with missing data from an unknown n-variable Ising model $\mathcal{D}(A,0)$, for $T \geq \frac{C\lambda^4}{\epsilon^4(1-p_{\max})^4}\exp\left(\lambda\left(\frac{2}{1-p_{\max}}+6\right)\right)\log(n/\delta)$ for large enough constant $C > 0$, Algorithm 4 returns $\bar{w}$ such that:*

$$\forall\, i \in [n-1],\ \left|\Pi^{n-1}_{\Delta\to B}(\bar{w})_i - A_{ni}\right| \leq \epsilon.$$

Proof We will focus on the missing-data model, same proof holds for random flipped-data model with a slightly different dependence on $p_{\max}$ that we highlight later in the proof. Observe that X^t is a new draw and is independent of $X^1, \ldots, X^{t-1}$ and therefore w^t, thus $G(w^t; X^t)$ is an unbiased estimator of $\nabla \widetilde{S}(w^t)$ (using Lemma 8) conditioned on $X^1, \ldots, X^{t-1}$. Applying Theorem 6 gives us that for $T \geq 16 \log n$ the output of Algorithm 4, $\bar{w}$, satisfies

$$\widetilde{S}(\bar{w}) - \widetilde{S}(w^*) \leq \frac{4\lambda}{(1-p_{\max})^2} \exp\left(\frac{\lambda}{1-p_{\max}}\right)\left(\sqrt{\frac{\log n}{T}} + \sqrt{\frac{2}{T}\log\frac{1}{\delta}}\right).$$

Recall that $w^* = \Pi_{B\to\Delta}^{n-1}(v^*)$ where v^* satisfies $v_i^* = A_{ni}$ for $i \in [n-1]$ such that $(i,n) \in E$ else 0. By definition of the mappings, we can see that $v^* = \Pi_{\Delta\to B}^{n-1}(w^*)$ and $\widetilde{S}(w^*) = S(v^*)$. Using Lemma 4, we have $\nabla S(v^*) = 0$, thus choosing $\bar{v} = \Pi_{\Delta\to B}^{n-1}(\bar{w})$ and applying the RSC property of S (see Lemma 5),

$$\begin{aligned}
||\bar{v} - v^*||_\infty^2 &\leq (1+\lambda)\exp(3\lambda)(S(\bar{v}) - S(v^*)) \\
&= (1+\lambda)\exp(3\lambda)(\widetilde{S}(\bar{w}) - \widetilde{S}(w^*)) \\
&\leq \frac{4(1+\lambda)\lambda}{(1-p_{\max})^2}\exp\left(\lambda\left(\frac{1}{1-p_{\max}}+3\right)\right)\left(\sqrt{\frac{\log n}{T}} + \sqrt{\frac{2}{T}\log\frac{1}{\delta}}\right).
\end{aligned}$$

Thus for given $T = O\left(\frac{\lambda^4}{\epsilon^4(1-p_{\max})^4}\exp\left(\lambda\left(\frac{2}{1-p_{\max}}+6\right)\right)\log(n/\delta)\right)$ we get the desired result. ■

Similarly for flipped data, we get:

Theorem 11 *For known constants $p_i < 0.5$, given samples with flipped data from an unknown n-variable Ising model $\mathcal{D}(A, 0)$, for $T \geq \frac{C\lambda^4}{\epsilon^4(1-2p_{\max})^4}\exp\left(\lambda\left(\frac{2}{1-2p_{\max}}+6\right)\right)\log(n/\delta)$ with large enough constant $C > 0$, Algorithm 4 returns $\bar{w}$ such that:*

$$\forall\, i \in [n-1],\ \left|\Pi_{\Delta\to B}^{n-1}(\bar{w})_i - A_{ni}\right| \leq \epsilon.$$

Setting $\epsilon = \beta/2$ in the above theorems where β is the smallest edge weight will recover the neighborhood of vertex n exactly. To recover the entire graph, we can run Algorithm 4 for each vertex using the same samples. Thus with probability $1-\delta$, for $p_{\max}$ being a constant using $T = O\left(\frac{\lambda^4}{\beta^4}\log(n^2/\delta)\exp(C\lambda)\right)$ samples we will recover the entire graph. The runtime for Algorithm 4 is $O(nT)$ as the unbiased estimator requires $O(n)$ time to compute. Thus, the overall runtime to recover the entire graph is $O(n^2T) = \widetilde{O}(n^2)$.

5. Extension to unknown p

In this section, we will show how to extend the analysis for the missing data model to the case when $p_i = p$ for all i and p is unknown. The main approach is to use fresh samples to estimate p with $\widehat{p}$ such that $p - \widehat{p} \approx \delta/\sqrt{Tn}$ using T/δ^2 samples where T corresponds to the number of iterations needed for the SMG algorithm. We can compute an empirical estimate of p since we observe ? when an entry is missing (it is unclear how to do this for the flipped data model). Also note that there is no dependence on n since each sample gives n estimates for p, one for each coordinate. Subsequently we will show that the distribution of T samples using p and $\widehat{p}$ are within total variation distance $O(\delta)$ for constant p. It follows that we can use $\widehat{p}$ in our SMG algorithm and obtain the same guarantees while losing a factor of $O(\delta)$ in the failure probability.

5.1. Estimating p

Prior to running SMG, we will draw m samples. Let $\widehat{p}$ be the fraction of ? observed in the m samples. Since each ? is i.i.d., with probability p, we have by Chernoff,

$$\Pr[|p - \widehat{p}| \geq \epsilon] \leq \exp\left(\frac{-mn\epsilon^2}{2p(1-p)}\right)$$

For $\epsilon = \frac{\delta}{\sqrt{Tn}}$, we have with probability $1-\delta$, $|p - \widehat{p}| \leq \frac{\delta}{\sqrt{Tn}}$ using $m = \widetilde{O}(T/\delta^2)$ samples.

5.2. Distribution closeness using $\hat{p}$

Here we will show that the distribution $\mathcal{D}$ over T samples with missing probability p is δ close to the distribution $\widehat{\mathcal{D}}$ over T samples with missing probability $\widehat{p}$. In order to do so, we will first show that $\mathcal{D}$ is equivalent to the following distribution $\mathcal{D}_{\mathsf{perm}}$:

1. Concatenate all T samples into a nT-dimensional vector and permute all coordinates.
2. Select first $c \sim \mathsf{Bin}(nT, p)$ coordinates and replace with ?.
3. Unpermute and split back to T samples.

Here $\mathsf{Bin}(n, p)$ denotes the binomial distribution over n runs with success probability p. It is not hard to see that $\mathcal{D}_{\mathsf{perm}}$ is equivalent to $\mathcal{D}$. Similarly the distribution $\widehat{\mathcal{D}}_{\mathsf{perm}}$, analogous to $\mathcal{D}_{\mathsf{perm}}$ with p replaced by $\hat{p}$ is equivalent to $\widehat{\mathcal{D}}$. Thus, we have

$$d_{\mathsf{TV}}(\mathcal{D}, \widehat{\mathcal{D}}) = d_{\mathsf{TV}}(\mathcal{D}_{\mathsf{perm}}, \widehat{\mathcal{D}}_{\mathsf{perm}}) = d_{\mathsf{TV}}(\mathsf{Bin}(nT, p), \mathsf{Bin}(nT, \hat{p})).$$

The above follows as $\mathcal{D}_{\mathsf{perm}}$ and $\widehat{\mathcal{D}}_{\mathsf{perm}}$ differ only in Step 2, which depends on only closeness of the binomial distribution.

We will use the following lemma to bound the closeness of binomial distributions.

Lemma 12 (Roos (2001)) *For $0 < p < 1$ and $0 < \delta < 1 - p$,*

$$d_{\mathsf{TV}}(\mathsf{Bin}(n, p), \mathsf{Bin}(n, p+\delta)) \leq \frac{\sqrt{e}}{2} \frac{\theta(\delta)}{(1-\theta(\delta))^2} \quad \textit{where} \quad \theta(\delta) := \delta\sqrt{\frac{n+2}{2p(1-p)}}.$$

The above gives us,

$$d_{\mathsf{TV}}(\mathcal{D}, \widehat{\mathcal{D}}) = d_{\mathsf{TV}}(\mathsf{Bin}(nT, p), \mathsf{Bin}(nT, \hat{p})) \leq O(\delta).$$

This implies that with probability $1 - O(\delta)$, the draw from $\mathcal{D}$ will be identical to that of $\widehat{\mathcal{D}}$. Thus, Algorithm 4 would give the desired result with estimate $\hat{p}$.

6. Conclusions and Open Problems

Our result highlights the importance of choosing the right surrogate loss to obtain noise-tolerant algorithms for learning the structure of Ising models. It would be interesting to know if other regression-based algorithms (e.g., the Sparsitron due to Klivans and Meka (2017)) can learn with independent failures. Other open problems include learning with missing data for distinct and *unknown* error rates $p_1, \ldots, p_n$ (we can only handle the unknown error-rate case when all p_i are equal) and improving the dependence on p in the sample complexity bound.

Acknowledgments

We thank Shanshan Wu for useful discussions.

References

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

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

Guy Bresler, David Gamarnik, and Devavrat Shah. Hardness of parameter estimation in graphical models. In *Advances in Neural Information Processing Systems*, pages 1062–1070, 2014.

Yuxin Chen. Learning sparse ising models with missing data.

Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. 2010.

C Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

Sanjoy Dasgupta. Learning polytrees. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 134–141. Morgan Kaufmann Publishers Inc., 1999.

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.

Ariel Jaimovich, Gal Elidan, Hanah Margalit, and Nir Friedman. Towards an integrated protein–protein interaction network: A relational markov network approach. *Journal of Computational Biology*, 13(2):145–164, 2006.

Sham Kakade, Dean Foster, and Eyal Even-Dar. (exponentiated) stochastic gradient descent for l1 constrained problems, 2008.

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

Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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.

Erik M Lindgren, Vatsal Shah, Yanyao Shen, Alexandros G. Dimakis, and Adam Klivans. On robust learning of ising models. 2018.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Richard Bonneau, Yukun Chen, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796, 2012.

Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using ?1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Bero Roos. Binomial approximation to the poisson binomial distribution: The krawtchouk expansion. *Theory of Probability & Its Applications*, 45(2):258–272, 2001.

Devavrat Shah and Dogyoon Song. Learning mixture model with missing values and its application to rankings. *arXiv preprint arXiv:1812.11917*, 2018.

Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *Journal of the ACM (JACM)*, 62(2):13, 2015.

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.

Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. *arXiv preprint arXiv:1810.11905*, 2018.

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.

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

Appendix A. Omitted Proofs

A.1. Proof of Lemma 4

Computing the gradient, we have

$$\begin{aligned}\nabla S(v^*)_i &= -\mathbb{E}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j|(n,j)\in E} A_{nj}Z_nZ_j\right)Z_nZ_i\right] \\ &= -\frac{1}{\mathcal{Z}\sim\mathcal{D}}\sum_{z\in\{-1,1\}^n}\exp\left(-\sum_{j|(n,j)\in E}A_{nj}z_nz_j\right)\exp\left(\sum_{(i,j)\in E}A_{ij}z_iz_j\right)z_nz_i \\ &= -\frac{1}{\mathcal{Z}\sim\mathcal{D}}\left(\sum_{z_{-n}\in\{-1,1\}^{n-1}}\exp\left(\sum_{i,j\neq n,(i,j)\in E}A_{ij}z_iz_j\right)z_i\right)\left(\sum_{z_n\in\{-1,1\}}z_n\right)=0\end{aligned}$$

Since S is convex, gradient is 0 at v^* and v^* lies on the l_1-ball of radius λ by assumption, it is the global minimum.

A.2. Proof of Lemma 5

To prove the above property, we use Lemma 5 from Vuffray et al. (2016) to get for all v,

$$S(v) - S(v^*) \geq \nabla S(v^*)^T(v-v^*) + \frac{\exp(-\lambda)}{1+||v-v^*||_1}(v-v^*)^TH(v-v^*)$$

where H satisfies $H_{ij} = \mathbb{E}_{Z\sim\mathcal{D}}[Z_iZ_j]$ for $i,j \in [n-1]$. The above property in their paper was proved for the empirical loss, however the same proof goes through for the expected loss with H being the true covariance matrix instead of the empirical one.

Let $\Delta = v - v^*$, now to bound $\Delta^TH\Delta$, we give an alternate stronger bound which holds for any vector on the l_1-ball and does not require some implicit sparsity. Note that we have $\Delta^TH\Delta = \sum_{i,j=1}^{n-1}\Delta_i\Delta_j\mathbb{E}_{Z\sim\mathcal{D}}[Z_iZ_j] = \mathbb{E}_{Z\sim\mathcal{D}}\left[\left(\sum_{i=1}^{n-1}\Delta_iZ_i\right)^2\right]$. Now for any $j \neq n$, we can condition the expectation as follows:

$$\begin{aligned}&\Delta^TH\Delta \\ &= \sum_{z_{-n}\in\{-1,1\}^{n-1}}\Pr[Z_{-n}=z_{-n}]\left(\sum_{i=1}^{n-1}\Delta_iZ_i\right)^2 \\ &= \sum_{z_{-\{j,n\}}\in\{-1,1\}^{n-2}}\Pr[Z_j=1|Z_{-\{j,n\}}=z_{-\{j,n\}}]\Pr[Z_{-\{j,n\}}=z_{-\{j,n\}}]\left(\Delta_j+\sum_{i=1,i\neq j}^{n-1}\Delta_iz_i\right)^2 \\ &\quad+\sum_{z_{-\{j,n\}}\in\{-1,1\}^{n-2}}\Pr[Z_j=-1|Z_{-\{j,n\}}=z_{-\{j,n\}}]\Pr[Z_{-\{j,n\}}=z_{-\{j,n\}}]\left(-\Delta_j+\sum_{i=1,i\neq j}^{n-1}\Delta_iz_i\right)^2 \\ &\geq \frac{\exp(-2\lambda)}{2}\sum_{z_{-\{j,n\}}\in\{-1,1\}^{n-2}}\Pr[Z_{-\{j,n\}}=z_{-\{j,n\}}]\left(\left(\Delta_j+\sum_{i=1,i\neq j}^{n-1}\Delta_iz_i\right)^2+\left(-\Delta_j+\sum_{i=1,i\neq j}^{n-1}\Delta_iz_i\right)^2\right)\end{aligned}$$

$$\geq \frac{\exp(-2\lambda)}{2} \sum_{z_{-\{j,n\}} \in \{-1,1\}^{n-2}} \Pr[Z_{-\{j,n\}} = z_{-\{j,n\}}] \left(2\Delta_j^2 + 2 \left(\sum_{i=1, i\neq j}^{n-1} \Delta_i z_i \right)^2 \right)$$
$$\geq \exp(-2\lambda) v_j^2 \sum_{z_{-\{j,n\}} \in \{-1,1\}^{n-2}} \Pr[Z_{-\{j,n\}} = z_{-\{j,n\}}] = \exp(-2\lambda)\Delta_j^2.$$

In the above, the first inequality follows from observing that $\Pr[Z_j = 1 | Z_{-\{j,n\}} = z_{-\{j,n\}}] \geq \frac{\exp(-2\lambda)}{2}$ using Lemma 3. Thus we get, $\Delta^T H \Delta \geq \exp(-2\lambda)||\Delta||_\infty^2$ which gives us the result.

A.3. Proof of Theorem 6

Since c is convex, we have for all iterations $c(u^*) - c(u^t) \geq \nabla c(u^t)^T (u^* - u^t)$. Applying Jensen's inequality, we have,

$$c(\bar{u}) - c(u^*) \leq \frac{1}{T} \sum_{i=1}^{T} (c(u^t) - c(u^*)) \leq \frac{1}{T} \sum_{i=1}^{T} \nabla c(u^t)^T (u^t - u^*).$$

Since $\mathbb{E}[g^t | \mathcal{H}^{t-1}] = \nabla c(u^t)$, we have $M^t = \sum_{r=1}^{t} (g^r - \nabla c(u^r))^T (u^r - u^*)$ is a martingale with respect to the filtration $\mathcal{H}^{t-1}$. Also $\mathbb{E}[M^1] = 0$ and $|M^t - M^{t-1}| = (g^t - \nabla c(u^t))^T (u^t - u^*) \leq 4BW$. By Azuma Hoeffding bound, we have with probability $1 - \delta$,

$$\frac{1}{T} \sum_{i=1}^{T} (\nabla c(u^t) - g^t)^T (u^t - u^*) \leq 4BW \sqrt{\frac{2}{T} \log \frac{1}{\delta}}.$$

Using the above, we have

$$c(\bar{u}) - c(u^*) \leq \frac{1}{T} \sum_{i=1}^{T} (g^t)^T (u^t - u^*) + 4BW \sqrt{\frac{2}{T} \log \frac{1}{\delta}}.$$

Let us now bound the first term. Note that for $\eta \leq \frac{1}{8B}$ we have $|\eta g_i^t - \eta C^t| \leq 2\eta B \leq 1/4$. Thus, each update keeps the weights positive. Also observe that, $||u^{t+1}||_1 = ||u^t||_1$ by construction, implying that each u^t remains in the feasible set.

Consider the following:

$$\begin{aligned}
\mathsf{KL}\left(\frac{u^*}{W} \middle|\middle| \frac{u^t}{W}\right) - \mathsf{KL}\left(\frac{u^*}{W} \middle|\middle| \frac{u^{t+1}}{W}\right) &= \sum_{i=1}^{k} \frac{u_i^*}{W} \log \frac{u_i^{t+1}}{u_i^t} \\
&= \frac{1}{W} \sum_{i=1}^{k} u_i^* \log \left(1 - \eta g_i^t + \frac{\eta (g^t)^T u^t}{W}\right) \\
&\geq \frac{1}{W} \sum_{i=1}^{k} u_i^* \left(-\eta g_i^t + \frac{\eta (g^t)^T u^t}{W}\right) - \frac{1}{W} \sum_{i=1}^{k} u_i^* \left(-\eta g_i^t + \frac{\eta (g^t)^T u^t}{W}\right)^2 \\
&\geq \frac{\eta (g^t)^T (u^t - u^*)}{W} - 4\eta^2 B^2.
\end{aligned}$$

Here the first inequality follows from observing that $\log(1+x) \geq x - x^2$ for all x such that $x \geq -1/4$. The second inequality follows from observing that $\sum_{i=1}^{k} \frac{u_i^*}{W} (g^t)^T u^t = (g^t)^T u^t \sum_{i=1}^{k} \frac{u_i^*}{W} = (g^t)^T u^t$. Rearranging, we have,

$$\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T}(g^t)^T(u^t - u^*) &\leq \frac{W}{\eta T}\left(\mathsf{KL}\left(\frac{u^*}{W}\middle\|\frac{u^1}{W}\right) - \mathsf{KL}\left(\frac{u^*}{W}\middle\|\frac{u^{t+1}}{W}\right)\right) + 4\eta B^2 W \\
&\leq \frac{W\log k}{\eta T} + 4\eta B^2 W \\
&\leq BW\sqrt{\frac{\log k}{T}}
\end{aligned}$$

for $\eta = \frac{1}{2B}\sqrt{\frac{\log k}{T}}$. The first inequality follows from observing that $\mathsf{KL}\left(\frac{u^*}{W}\middle\|\frac{u^1}{W}\right) \leq \log k$. Observe that we needed $\eta \leq \frac{1}{8B}$, this implies that $T \geq 16\log k$. This gives us the desired result.

A.4. Proof of Lemma 9

Similar to the proof of 8, let $v = \Pi_{\Delta\to B}^{n-1}$ that is, $v_i = w_i - w_{n-1+i}$. Taking expectation of g_{miss} over $\mathcal{D}_{\mathsf{flip}}$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{flip}}}[f(w;X)_i] \\
&= \mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{flip}}}\left[X_nX_i \times \prod_{j\neq n}\sigma(p_j, -v_jX_n, X_j)\right] \\
&= \mathop{\mathbb{E}}_{\substack{Z\sim\mathcal{D}\\ C_n\sim\mathsf{Rad}(1-p_n)}}\left[C_nZ_n \mathop{\times}_{C_i\sim\mathsf{Rad}(1-p_i)} \mathbb{E}\left[C_iZ_i\sigma(p_i, -v_iC_nZ_n, C_iZ_i)\right] \times \right. \\
&\qquad \left. \prod_{j\neq i,n} \mathop{\mathbb{E}}_{C_j\sim\mathsf{Rad}(1-p_j)}\left[\sigma(p_j, -v_jC_nZ_n, C_jZ_j)\right]\right] \\
&= \mathop{\mathbb{E}}_{\substack{Z\sim\mathcal{D}\\ C_n\sim\mathsf{Rad}(1-p_n)}}\left[C_nZ_n \times \exp\left(-v_iC_nZ_nZ_i\right)Z_i \times \prod_{j\neq i,n}\exp\left(-v_jC_nZ_nZ_j\right)\right] \\
&= \mathop{\mathbb{E}}_{\substack{Z\sim\mathcal{D}\\ C_n\sim\mathsf{Rad}(1-p_n)}}\left[\exp\left(-\sum_{j=1}^{n-1}v_iX_nZ_j\right)C_nZ_nZ_i\right] \\
&= \mathop{\mathbb{E}}_{Z\sim\mathcal{D}}\left[\left((1-p_n)\exp\left(-\sum_{j=1}^{n-1}v_iZ_nZ_j\right) - p_n\exp\left(\sum_{j=1}^{n-1}v_iZ_nZ_j\right)\right)Z_nZ_i\right].
\end{aligned}$$

The above follows from observing the following two facts that follow from the example mentioned in the discussion: for all i 1) $\mathbb{E}_{C_i}[\sigma(p, A, C_iZ_i)] = \exp(AZ_i)$, 2) $\mathbb{E}_{C_i}[X_i\sigma(p, A, C_iZ_i)] = \exp(AZ_i)Z_i$ as long as A is independent of C_i.

Using the above, we have

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{flip}}}[g^i_{\mathsf{flip}}(w;X)] = -\frac{(1-p_n)\mathbb{E}_X[f(w;X)_i] + p_n\mathbb{E}_X[f(-w;X)_i]}{1-2p_n}$$

$$= -\mathop{\mathbb{E}}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j=1}^{n-1} v_i Z_n Z_j\right) Z_n Z_i\right] = \nabla\widetilde{S}(w)_i.$$

The last thing we need is that the above estimator has bounded l_∞ norm for all $w \in \Delta(\lambda, 2n-1)$ ($v \in B(\lambda, n-1)$). We have

$$\begin{aligned}
|\sigma(p,a,X)| &= \frac{|(1-p)\exp(aX) - p\exp(-aX)|}{1-2p} \\
&= \frac{\left|(1-p)\left(1+\sum_{i=1}^{\infty}\frac{(aX)^i}{i!}\right) - p\left(1+\sum_{i=1}^{\infty}(-1)^i\frac{(aX)^i}{i!}\right)\right|}{1-2p} \\
&= \left|1 + \sum_{i=1}^{\infty}\left(\frac{(aX)^{2i}}{(2i)!} + \frac{(aX)^{2i-1}}{(1-2p)(2i-1)!}\right)\right| \\
&\leq 1 + \sum_{i=1}^{\infty}\frac{|aX|^i}{(1-2p)i!} \leq \exp\left(\frac{|aX|}{1-2p}\right).
\end{aligned}$$

Let $p_{\max} := \max_i p_i$, then using the fact that $|X| = 1$ we get

$$\begin{aligned}
|f(w;X)^i| &\leq \frac{\exp(|v_i|)}{1-2p_i}\prod_{j\neq i,n}\exp\left(\frac{|w_j - w_{n-1+j}|}{1-2p_j}\right) \\
&= \frac{1}{1-2p_{\max}}\exp\left(\frac{\lambda}{1-2p_{\max}}\right).
\end{aligned}$$

Here the inequality follows from observing that $\sum_{i=1}^{n-1}|v_i| = ||v||_1 \leq \lambda$. Using above, we get the final bound on $||G_{\mathsf{flip}}(w;X)||_\infty \leq \frac{1}{(1-p_{\max})^2}\exp\left(\frac{\lambda}{1-p_{\max}}\right)$.

Appendix B. Ising Models with Non-zero Mean Field

In this section, we will extend the analysis to the setting $\theta \neq 0$. Similar to the case of zero-mean field, consider the neighborhood of a fixed vertex (WLOG we choose n, we can do the same analysis for any $v \in V$). We will work with the following modified ISO:

$$\begin{aligned}
\min_{v\in\mathbb{R}^n} \quad & \underline{S}(v) := \mathbb{E}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j=1}^{n-1} v_j Z_n Z_j - v_n Z_n\right)\right] \\
\text{subject to} \quad & ||v||_1 \leq \lambda_{\neq 0}.
\end{aligned}$$

Define $\underline{v}^* \in \mathbb{R}^{n-1}$ such that $\underline{v}^*_n = \theta_n$ and for $i \in [n-1]$, $\underline{v}^*_i = A_{ni}$ if $(i,n) \in E$ else 0. It is not hard to see that the above loss is convex and following the same line of proof of Lemma 4, one can also show that $\nabla\underline{S}(\underline{v}^*) = 0$ making it an optimal solution. Further, one can prove a close to RSC property for the given objective.

Lemma 13 *For all v,*

$$\underline{S}(v) - \underline{S}(\underline{v}^*) \geq \nabla\underline{S}(\underline{v}^*)^T(v - \underline{v}^*) + \frac{\exp(-3\underline{\lambda})}{1+\lambda}||(v-\underline{v}^*)_{-n}||_\infty.$$

Proof Consider the following transformation for the input variables of the Ising model, $Z \to (Z, 1)$. We can then use Lemma 5 from Vuffray et al. (2016) to get for all v,

$$\underline{S}(v) - \underline{S}(v^*) \geq \nabla \underline{S}(v^*)^T (v - v^*) + \frac{\exp(-\lambda)}{1 + ||v - v^*||_1} (v - v^*)^T \underline{H} (v - v^*)$$

where $\underline{H}$ is as follows:

$$\underline{H}_{ij} = \begin{cases} 1 & \text{for } i = j = n \\ \mathbb{E}_{Z \sim \mathcal{D}}[Z_i] & \text{for } i \in [n-1], j = n \\ \mathbb{E}_{Z \sim \mathcal{D}}[Z_j] & \text{for } i = n, j \in [n-1] \\ \mathbb{E}_{Z \sim \mathcal{D}}[Z_i Z_j] & \text{otherwise.} \end{cases}$$

We will now bound $v^T \underline{H} v$, we have, $v^T H v = \mathbb{E}_{Z \sim \mathcal{D}} \left[\left(\sum_{i=1}^{n-1} v_i Z_i + v_n \right)^2 \right]$. Similar to the proof of Lemma 5, for all $j \in [n-1]$,

$$\begin{aligned}
& v^T \underline{H} v \\
&= \sum_{z_{-n} \in \{-1,1\}^{n-1}} \Pr[Z_{-n} = z_{-n}] \left(\sum_{i=1}^{n-1} v_i Z_i + v_n \right)^2 \\
&= \sum_{z_{-\{j,n\}} \in \{-1,1\}^{n-2}} \Pr[Z_j = 1 | Z_{-\{j,n\}} = z_{-\{j,n\}}] \Pr[Z_{-\{j,n\}} = z_{-\{j,n\}}] \left(v_j + \sum_{i=1, i \neq j}^{n-1} v_i z_i + v_n \right)^2 \\
&\quad + \sum_{z_{-\{j,n\}} \in \{-1,1\}^{n-2}} \Pr[Z_j = -1 | Z_{-\{j,n\}} = z_{-\{j,n\}}] \Pr[Z_{-\{j,n\}} = z_{-\{j,n\}}] \left(-v_j + \sum_{i=1, i \neq j}^{n-1} v_i z_i + v_n \right)^2 \\
&\geq \frac{\exp(-2\underline{\lambda})}{2} \sum_{z_{-\{j,n\}} \in \{-1,1\}^{n-2}} \Pr[Z_{-\{j,n\}} = z_{-\{j,n\}}] \left[2v_j^2 + 2 \left(\sum_{i=1, i \neq j}^{n-1} v_i z_i + v_n \right)^2 \right] \\
&\geq \exp(-2\underline{\lambda}) v_j^2.
\end{aligned}$$

Thus, we have that $v^T \underline{H} v \geq \exp(-2\underline{\lambda}) ||v_{-n}||_\infty^2$. Substituting this in the previous equation gives us the desired result. ■

Similar to the zero mean-field case, we can transform the problem from $B(\lambda, n)$ to $\Delta(\lambda, 2n+1)$ to get the following optimization problem:

$$\begin{aligned}
\min_{w \in \mathbb{R}^{2n+1}} \quad & \widetilde{\underline{S}}(w) := \underline{S}(\Pi^n_{\Delta \to B}(w)) = \mathbb{E}_{Z \sim \mathcal{D}} \left[\exp \left(- \sum_{j=1}^{n-1} (w_j - w_{n+j}) Z_n Z_j - (w_n - w_{2n}) Z_n \right) \right] \\
\text{subject to} \quad & ||w||_1 = \lambda, w \geq 0.
\end{aligned}$$

Note that the above loss is also convex and similar to Lemma 4, we can show that $\underline{w}^* = \Pi^n_{B \to \Delta}(\underline{v}^*)$ for $\underline{v}^*$ defined previously, is an optimal solution of the above optimization (by showing a zero

gradient value). Thus for running SMG, we only need to show the existence of an unbiased estimator. Observe that the gradient of $\widetilde{\underline{S}}(w)$ with respect to w can be computed as follows, we have $\nabla\widetilde{\underline{S}}(w)_{2n+1} = 0, \forall i \in [n-1]$,

$$\nabla\widetilde{\underline{S}}(w)_i = -\nabla\widetilde{\underline{S}}(w)_{i+n} = -\mathbb{E}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j=1}^{n-1}(w_j - w_{n+j})Z_nZ_j - (w_n - w_{2n})Z_n\right)Z_nZ_i\right]$$

and

$$\nabla\widetilde{\underline{S}}(w)_n = -\nabla\widetilde{\underline{S}}(w)_{2n} = -\mathbb{E}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j=1}^{n-1}(w_j - w_{n+j})Z_nZ_j - (w_n - w_{2n})Z_n\right)Z_n\right].$$

Here we will focus on only the missing-data model. Similar ideas extend to the flipped-data model also. Consider the following estimators, $\forall i \in [n-1]$

$$\underline{g}^i_{\mathsf{miss}}(w;X) = -\frac{\exp\left(-(w_n - w_{2n})X_n\right)X_n}{1-p_n} \times \frac{\exp\left(-(w_i - w_{n+i})X_nX_i\right)X_i}{1-p_i} \times \prod_{j\neq i,n}\frac{\exp\left(-(w_j - w_{n+j})X_nX_j\right) - p_j}{1-p_j}$$

and

$$\underline{g}^n_{\mathsf{miss}}(w;X) = -\frac{\exp\left(-(w_n - w_{2n})X_n\right)X_n}{1-p_n} \times \prod_{j\neq n}\frac{\exp\left(-(w_j - w_{n-1+j})X_nX_j\right) - p_j}{1-p_j}.$$

We will show that $G(w;X) = \sum_{i=1}^n \underline{g}^i_{\mathsf{miss}}(e^i - e^{n+i})$ for $e^i \in \mathbb{R}^{2n+1}$ is an indicator vector for coordinate i, is an unbiased estimator of the gradient of $\widetilde{\underline{S}}$. Let $v = \Pi^n_{\Delta\to B}(w)$. Taking expectation of g_{miss} over $\mathcal{D}_{\mathsf{miss}}$, we have for all $i \in [n-1]$

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{miss}}}[\underline{g}^i_{\mathsf{miss}}(w;X)] \\
&= -\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{miss}}}\left[\frac{\exp\left(-v_nX_n\right)X_n}{1-p_n} \times \frac{\exp\left(-v_iX_nX_i\right)X_i}{1-p_i} \times \prod_{j\neq i,n}\frac{\exp\left(-v_jX_nX_j\right) - p_i}{1-p_i}\right] \\
&= -\mathop{\mathbb{E}}_{\substack{Z\sim\mathcal{D}\\ C_n\sim\mathsf{Ber}(1-p_n)}}\left[\frac{\exp\left(-v_nC_nZ_n\right)C_nZ_n}{1-p_n} \times \mathop{\mathbb{E}}_{C_i\sim\mathsf{Ber}(1-p_i)}\left[\frac{\exp\left(-v_iC_nZ_nC_iZ_i\right)C_iZ_i}{1-p_i}\right] \times \right. \\
&\qquad \left. \prod_{j\neq i,n}\mathop{\mathbb{E}}_{C_j\sim\mathsf{Ber}(1-p_j)}\left[\frac{\exp\left(-v_jC_nZ_nC_jZ_j\right) - p_j}{1-p_j}\right]\right] \\
&= -\mathop{\mathbb{E}}_{\substack{Z\sim\mathcal{D}\\ C_n\sim\mathsf{Ber}(1-p_n)}}\left[\frac{\exp\left(-\sum_{j\neq n}v_jC_nZ_nZ_j - v_nC_nZ_n\right)Z_iC_nZ_n}{1-p_n}\right] \\
&= -\mathop{\mathbb{E}}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j\neq n}v_jZ_nZ_j - v_nZ_n\right)Z_iZ_n\right] = \nabla\widetilde{\underline{S}}(w)_i.
\end{aligned}$$

Similarly,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{miss}}}[\underline{g}^n_{\mathsf{miss}}(w;X)] \\
&= -\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{miss}}}\left[\frac{\exp\left(-v_nX_n\right)X_n}{1-p_n}\prod_{j\neq n}\frac{\exp\left(-v_jX_nX_j\right)-p_j}{1-p_j}\right] \\
&= -\mathop{\mathbb{E}}_{\substack{Z\sim\mathcal{D}\\ C_n\sim\mathsf{Ber}(1-p_n)}}\left[\frac{\exp\left(-v_nC_nZ_n\right)C_nZ_n}{1-p_n}\prod_{j\neq n}\mathop{\mathbb{E}}_{C_j\sim\mathsf{Ber}(1-p_j)}\left[\frac{\exp\left(-v_jC_nZ_nC_jZ_j\right)-p_j}{1-p_j}\right]\right] \\
&= -\mathop{\mathbb{E}}_{\substack{Z\sim\mathcal{D}\\ C_n\sim\mathsf{Ber}(1-p_n)}}\left[\frac{\exp\left(-\sum_{j\neq n}v_jC_nZ_nZ_j-v_nC_nZ_n\right)C_nZ_n}{1-p_n}\right] \\
&= -\mathop{\mathbb{E}}_{Z\sim\mathcal{D}}\left[\exp\left(-\sum_{j\neq n}v_jZ_nZ_j-v_nZ_n\right)Z_n\right] = \nabla\widetilde{\underline{S}}(w)_n.
\end{aligned}
$$

This gives us that $\mathop{\mathbb{E}}_{X\sim\mathcal{D}_{\mathsf{miss}}}[G_{\mathsf{miss}}(w;X)] = \nabla\widetilde{\underline{S}}(w)$. Lastly we can see that the norm bound of the estimator follows the same proof as the zero-mean field case and we get, $||\underline{G}_{\mathsf{miss}}(w;X)||_\infty \leq \frac{1}{(1-p_{\max})^2}\exp\left(\frac{\lambda}{1-p_{\max}}\right)$ for all $w\in\Delta(\lambda,2n+1)$.

Now we can straightforwardly apply Theorem 10 to get the following:

Theorem 14 *For known constants p_i, given samples with missing data from an unknown n-variable Ising model $\mathcal{D}(A,\theta)$, for $T\geq\lambda^4\log(n/\delta)\exp(C\lambda)\epsilon^{-4}$ with large enough constant $C>0$, Algorithm 4 returns $\bar{w}$ such that:*

$$\forall\, i\in[n-1],\ \left|\Pi^{n-1}_{\Delta\to B}(\bar{w})_i - A_{ni}\right|\leq\epsilon.$$