On Learning Ising Models under Huber's Contamination Model

Adarsh Prasad[*] [†] **Vishwak Srinivasan**[*] [†] **Sivaraman Balakrishnan** [‡] [†] **Pradeep Ravikumar** [†]

adarshp@andrew.cmu.edu, vishwaks@cs.cmu.edu
siva@stat.cmu.edu, pradeepr@cs.cmu.edu

[†] Machine Learning Department
[‡] Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We study the problem of learning Ising models in a setting where some of the samples from the underlying distribution can be arbitrarily corrupted. In such a setup, we aim to design statistically optimal estimators in a high-dimensional scaling in which the number of nodes p, the number of edges k and the maximal node degree d are allowed to increase to infinity as a function of the sample size n. Our analysis is based on exploiting moments of the underlying distribution, coupled with novel reductions to univariate estimation. Our proposed estimators achieve an optimal dimension independent dependence on the fraction of corrupted data in the contaminated setting, while also simultaneously achieving high-probability error guarantees with optimal sample-complexity. We corroborate our theoretical results by simulations.

1 Introduction

Undirected graphical models (also known as Markov random fields (MRFs)) have gained significant attention as a tool for discovering and visualizing dependencies among covariates in multivariate data. Graphical models provide compact and structured representations of the joint distribution of multiple random variables using graphs that represent conditional independences between the individual random variables. They are used in domains as varied as natural language processing[37], image processing [9, 24, 26], spatial statistics [43] and computational biology [23], among others. Given samples drawn from the distribution, a key problem of interest is to recover the underlying dependencies represented by the graph. A slew of recent results [39, 42, 44] have shown that it is possible to learn such models even in domains and settings where the number of samples is potentially smaller than the number of variables. These results however make the common assumption that the sample data is clean, and have no corruptions. However, modern data sets that arise in various branches of science and engineering are no longer carefully curated. They are often collected in a decentralized and distributed fashion, and consequently are plagued with the complexities of outliers, and even adversarial manipulations.

Huber [27] proposed the ϵ-contamination model as a framework to study such datasets with potentially arbitrary corruptions. In this setting, instead of observing samples directly from the true distribution $\mathbb{P}^{\star}$, we observe samples drawn from $\mathbb{P}_{\epsilon}$, which for an arbitrary distribution Q is defined as a mixture

[*] Equal Contribution

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

model,

$$\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}^\star + \epsilon Q. \tag{1}$$

Then, given n samples from $\mathbb{P}_\epsilon$, the goal is to recover functionals of $\mathbb{P}^\star$. There has been a lot of classical work on estimators for the ϵ-contamination model setting that largely trade off computational versus statistical efficiency (see [28] and references therein). Moreover, there has been substantial progress [3, 7, 15, 16, 18, 31, 34, 41] on designing provably robust estimators which are computationally tractable while achieving near-optimal contamination dependence (*i.e.* dependence on the fraction of outliers ϵ). However, to the best of our knowledge, there are no known results for learning general graphical models robustly.

1.1 Related Work

In this work, we focus on the specific undirected graphical model sub-class of Ising models [29]. There has been a lot of work for learning Ising models in the uncontaminated setting dating back to the classical work of Chow and Liu [8]. Csiszár and Talata [10] discuss pseudo likelihood based approaches for estimating the neighborhood at a given node in MRFs. Subsequently, a simple search based method is described in [6] with provable guarantees. Later, Ravikumar et al. [42] showed that under an incoherence assumption, node-wise (regularized) estimators provably recover the correct dependency graph with a small number of samples. Recently, there has been a flurry of work [5, 30, 36, 47, 49] to get computationally efficient estimators which recover the true graph structure without the incoherence assumption, including extensions to identity and independence testing [12]. However, all the aforementioned results are in the uncontaminated setting. Recently, Lindgren et al. [35] derived preliminary results for learning Ising models robustly. However, their upper and lower bounds do not match. Moreover, their analysis primarily focuses on the robustness of the Sparsitron algorithm in [30], and they do not explore the effect of the underlying graph and correlation structures comprehensively.

Contributions. In this work, we give the *first* statistically optimal estimator for learning Ising models under the ϵ-contamination model. Our estimators achieve a dimension-independent asymptotic error as a function of the fraction of outliers ϵ, while simultaneously achieving high probability deviation bounds. As an important special case of our results, we also close known sample complexity gaps in the uncontaminated setting for some classes of Ising models. We finally corroborate our theoretical findings with simulation studies.

1.2 Background and Problem Setup

We begin with some background on Ising models and then provide the precise formulation of the problem. We follow the notation of Santhanam and Wainwright [45] very closely.

Consider an undirected graph $G = (V, E)$ defined over a set of vertices $V = \{1, 2, \ldots, p\}$ with edges $E \subset \{(s,t) : s, t \in V, s \neq t\}$. The neighborhood of any node $s \in V$ is the subset $\mathcal{N}(s) \subset V$ given by $\mathcal{N}(s) \overset{\text{def}}{=} \{t | (s,t) \in E\}$, and the degree of any vertex s is given by $d_s = |\mathcal{N}(s)|$. Then, the degree of a graph $d = \max_s d_s$ is the maximum vertex degree, and $k = |E|$ is the total number of edges. We obtain an MRF by associating a random variable X_v at each vertex $v \in V$, and then considering a joint distribution $\mathbb{P}$ over the random vector $(X_1, \ldots, X_p)$. An Ising model is a special instantiation of an MRF where each random variable X_s take values in $\{-1, +1\}$, and the joint probability mass function is given by:

$$\mathbb{P}_\theta(x_1, \ldots, x_p) \propto \exp\left(\sum_{1 \le s < t \le p} \theta_{st} x_s x_t\right), \tag{2}$$

where we view θ as the parameter vector of the distribution. Note that $\theta \in \mathbb{R}^{p \times p}$ is such that $\theta_{ij} = 0 \Leftrightarrow (i,j) \notin E$ and $\theta = \theta^T$.

Graph Classes. In this work, we consider two classes of Ising models (2) based on the conditions imposed on the edge set:

1. $\mathcal{G}_{p,d}$: the collection of graphs G with p vertices such that each vertex has at most d neighbors for some $d \geq 1$, and

2. $\mathcal{G}_{p,k}$: the collection of graphs G with p vertices such that the total number of edges in the graph is at most k for some $k \geq 1$.

In addition to these structural properties, we also consider some subclasses based on the parameters of the Ising model. We define the *model width* as:

$$\omega^*(\theta(G)) \overset{\text{def}}{=} \max_{u \in V} \sum_{v \in V} |\theta_{uv}|.$$

It is well-known (see for instance [45]) that estimation in Ising models becomes harder with increasing value of edge parameters, since, large values of edge parameters may hide the contributions of other edges. Similarly, we define the *minimum edge weight* as:

$$\lambda^*(\theta(G)) \overset{\text{def}}{=} \min_{(s,t) \in E} |\theta_{st}| \, .$$

With these structural and parameter properties in place, we define the classes of Ising models that we will be studying in the rest of the paper. Given a pair of positive numbers (λ, ω):

1. $\mathcal{G}_{p,d}(\lambda, \omega)$: the set of all Ising models defined over a graphs G with p vertices, with each vertex having degree at most d and parameters satisfying
$$\lambda^*(\theta(G)) \geq \lambda \ \text{ and } \ \omega^*(\theta(G)) \leq \omega.$$
2. $\mathcal{G}_{p,k}(\lambda, \omega)$: the set of all Ising models defined over a graphs G with p vertices, with total number of edges at most k and parameters satisfying
$$\lambda^*(\theta(G)) \geq \lambda \ \text{ and } \ \omega^*(\theta(G)) \leq \omega.$$

Furthermore, we work in the **high temperature regime** where we assume that the model width bound $\omega^*(\theta(G)) \leq 1 - \alpha$ for some $\alpha > 0$. Note that this assumption implies the Dobrushin condition [19], which in case of Ising models is given by

$$\max_{u \in V} \sum_{v \in V} \tanh(|\theta_{uv}|) \leq 1 - \alpha, \qquad \alpha \in (0, 1). \tag{3}$$

While this may seem restrictive, this assumption is widely popular for studying Ising models, for example, see related works in statistical physics [20, 46], mixing times of Glauber dynamics [13, 32], correlation decay [33] and more recently in estimation and testing problems [11, 12].

Notation: Given a matrix M of dimensions $l \times m$, we will denote the i^{th} row of matrix by M_i or $M(i)$ and the $(i, j)^{th}$ element by M_{ij} or $M(i, j)$. M_{-i} or $M(-i)$ denotes the sub-matrix formed by all rows except i, and analogously $M_{:,-j}$ or $M(:, -j)$ denotes the sub-matrix formed by all columns except j. Given a vector v, $\|v\|_p = \sqrt[p]{\sum_i |v_i|^p}$ denotes its ℓ_p-norm, and its ℓ_∞-norm is given by $\|v\|_{\max} = \max_i |v_i|$. For a matrix M, $\|M\|_{p,q}$ denotes the mixed $\ell_{p,q}$-norm, which is the q-norm of the collection of p-norms of the rows of M. We also use the shorthand $[d] = \{1, 2, \ldots, d\}$. We denote the total variation (TV) distance between two discrete distributions p, q with support $\mathcal{X}$ by $d_{\text{TV}}(p, q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)|$.

2 Information-theoretic bounds for the ϵ-contamination model

Recall that in the ϵ-contamination model (1), we observe n samples from $\mathbb{P}_\epsilon = (1 - \epsilon)\mathbb{P}^\star + \epsilon Q$. In this model, even in the asymptotic setting as $n \to \infty$, we cannot expect to recover the true parameters exactly. To see this, suppose that $\mathbb{P}_1^\star, \mathbb{P}_2^\star$ are such that there exist two distributions Q_1 and Q_2 such that

$$\mathbb{P}_\epsilon = (1 - \epsilon)\mathbb{P}_1^\star + \epsilon Q_1 = (1 - \epsilon)\mathbb{P}_2^\star + \epsilon Q_2,$$

then, we cannot hope to distinguish between the two distributions. It is easy to show (see [17]) that the above condition is equivalent to assuming that $d_{\text{TV}}(\mathbb{P}_1^\star, \mathbb{P}_2^\star) = \frac{\epsilon}{1-\epsilon}$. Thus, for any given contaminated distribution $\mathbb{P}_\epsilon$, there is a set of possible uncontaminated distributions (including the ground truth uncontaminated distribution among others) within a ball of some fixed radius with respect to the TV distance, any of which could give rise to the given contaminated distribution $\mathbb{P}_\epsilon$. Thus, when estimating the uncontaminated distribution with respect to some loss function, in the worst case we could incur loss corresponding to the farthest pair of distributions in the ball of some fixed radius with respect to TV distance. This is captured by the geometric notion of modulus of continuity [22], which can then be used to derive sharp bounds on estimation in such a setting:

Definition 1 (TV modulus of continuity)**.** *Given a loss function $L : \Theta \times \Theta \to \mathbb{R}^+$ defined over the parameter space Θ, a class of distributions $\mathcal{D}$, a functional $f : \mathcal{D} \to \Theta$ and a proximity parameter ϵ, the modulus of continuity $\omega(f, \mathcal{D}, L, \epsilon)$ is defined as*

$$\omega(f, \mathcal{D}, L, \epsilon) \stackrel{\text{def}}{=} \sup_{\substack{\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{D} \\ d_{\mathrm{TV}}(\mathbb{P}_1, \mathbb{P}_2) \leq \epsilon}} L(f(\mathbb{P}_1), f(\mathbb{P}_2)). \tag{4}$$

Intuitively, this quantity controls how far the functionals of two distributions can be, subject to the constraint that the TV distance between them is ϵ. Note that for general Ising models, there do not exist *any* results that directly relate the total variation distance to the difference in parameters *i.e.* which study the TV modulus of continuity for the parameters of an Ising model.

A key contribution of our work is to establish sharp upper bounds on the TV modulus of continuity for parameter error in the high temperature regime. The loss function is considered to be the $\ell_{2,\infty}$ norm *i.e.* for matrices $x, y \in \mathbb{R}^{p \times p}$, $L(x, y) = \max_i \|x_i - y_i\|_2$.

Theorem 1. *Consider two Ising models defined over two graphs $G^{(1)}$ and $G^{(2)}$ with p vertices with parameters $\theta^{(1)}$ and $\theta^{(2)}$ respectively, each of which satisfy the high temperature condition (3) with constant α. If $d_{TV}\left(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}\right) \leq \epsilon$, then we have that:*

$$\|\theta^{(1)}(i) - \theta^{(2)}(i)\|_2 \lesssim {}^{1}\epsilon\sqrt{C_1(\alpha)\log\left(\frac{2}{\epsilon}\right)} \text{ for all } i \in [p],$$

where $C_1(\alpha)$ is a constant depending on α.

Observe that Theorem 1 shows that the parameter error is *independent* of the dimension p, degree d and the number of edges k. Furthermore, it is also independent of the minimum edge weight λ. As expected, when $\epsilon \to 0$, we see that the parameters are equal providing an alternate route to showing that the parameters of an Ising model are identifiable in the high temperature setting. We also establish that the dependence on ϵ is tight upto logarithmic factors by providing a complementary lower bound – proofs of which are made available in the appendix (Sections C.1 and C.2).

Lemma 1. *There exists two Ising models satisfying the properties in Theorem 1 whose parameters $\theta^{(1)}$ and $\theta^{(2)}$ satisfy:*

$$\|\theta^{(1)}(i) - \theta^{(2)}(i)\|_2 \gtrsim \epsilon \text{ for all } i \in [p].$$

3 TV Projection Estimators

Recall the geometric picture of TV contamination discussed in the previous section: given the contaminated distribution, there is a set of possible uncontaminated distributions within a ball of some fixed radius with respect to TV. It is thus natural to consider the TV projection of the contaminated distribution onto the set of all possible uncontaminated distributions. These are also called *minimum distance estimators* and were proposed by Donoho and Liu [21], which we consider for our setting to learn Ising models robustly, leveraging our Theorem 1.

3.1 Population Robust Estimators for $\mathcal{G}_p$

Let us first consider the population setting *i.e.*, in which we have distribution access to the contaminated distribution $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}_{\theta^\star} + \epsilon Q$, where $\mathbb{P}_{\theta^\star} \in \mathcal{G}_p(\lambda, \omega)$ [2]. In this setting, we use the minimum distance estimator [21] to construct robust estimators. In particular, let $\mathbb{P}_{\widehat{\theta}_{\mathrm{MDE}}}$ be the minimum distance estimate defined as

$$\mathbb{P}_{\widehat{\theta}_{\mathrm{MDE}}} = \operatorname*{argmin}_{\mathbb{P}_\theta \in \mathcal{G}_p} d_{\mathrm{TV}}(\mathbb{P}_\theta, \mathbb{P}_\epsilon). \tag{5}$$

This estimator is effectively the TV projection of the contaminated distribution onto the set of all Ising model distributions whose underlying graph lies in $\mathcal{G}_p$.

[1] Here and throughout our paper we use the notation $\lesssim$ to denote an inequality with universal constants dropped for conciseness.

[2] We define the class $\mathcal{G}_p(\lambda, \omega)$ as the set of Ising models defined over p vertices with minimum edge weight λ and model width ω

Noting that $d_{\mathrm{TV}}(\mathbb{P}_{\theta^\star}, \mathbb{P}_{\widehat{\theta}_{\mathrm{MDE}}}) \leq \epsilon$, by an application of the triangle inequality we have that $d_{\mathrm{TV}}(\mathbb{P}_{\theta^\star}, \mathbb{P}_{\widehat{\theta}_{\mathrm{MDE}}}) \leq 2\epsilon$. Combining this with Theorem 1, we get that,

$$\|\widehat{\theta}_{\mathrm{MDE}}(i) - \theta^\star(i)\|_2 \lesssim \epsilon\sqrt{C(\alpha)\log\left(\frac{2}{\epsilon}\right)} \text{ for all } i \in [p].$$

Corollary 1. *Let $\mathbb{P}_{\widehat{\theta}_{\mathrm{MDE},\lambda}}$ be the TV projection of the contaminated distribution $\mathbb{P}_\epsilon$ onto the class of Ising models $\mathcal{G}_{p,d}$ with minimum edge weight at least λ. Define the edge set of $\mathbb{P}_{\widehat{\theta}_{\mathrm{MDE}}}$ as $E(\widehat{\theta}_{\mathrm{MDE},\lambda}) = \{(i,j) : |\widehat{\theta}_{\mathrm{MDE},\lambda}(i,j)| > \frac{\lambda}{2}\}$. When $\epsilon\sqrt{C(\alpha)\log\left(\frac{2}{\epsilon}\right)} \leq \frac{\lambda}{2C_1}$, where C_1 is a universal constant, the edge sets of $\mathbb{P}_{\widehat{\theta}_{\mathrm{MDE},\lambda}}$ and $\mathbb{P}_{\theta^\star}$ coincide i.e.,*

$$E(\widehat{\theta}_{\mathrm{MDE},\lambda}) = E(\theta^\star).$$

Observe that this result is interesting and surprising, because one would generally not expect to be able to recover the true edge $E(\theta^\star)$ under contamination. Additionally, as mentioned earlier, there is no dependence on p, d or k, which means that the irrespective of the size of graph, if the minimum edge weight is sufficiently large or the level of contamination is sufficiently small, we would be able to recover the true edge set in the infinite sample limit.

3.2 Empirical Robust Estimators for $\mathcal{G}_{p,k}$

The minimum distance estimator is not suitable for non-asymptotic settings since we do not have access to the population contaminated distribution, but only to its discrete empirical counterpart, obtained via samples from the contaminated distribution. It would thus be ideal if there were an approximation to the TV distance that is amenable to projections of discrete distributions, and that preserves the optimality properties of the full TV projections.

Remarkably, Yatracos [50] proposed just such an approximation to TV projections. Consider a class of distributions $\mathcal{P}$. It is known that $d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$, where the supremum is over all possible measurable sets $A \subseteq \mathrm{supp}(P)$. While uniform convergence fails over all sets, Yatracos [50] showed that we can consider a much smaller collection of clevely chosen sets. In particular, Yatracos [50] suggested approximating the TV distance between distribution $P, Q \in \mathcal{P}$ as

$$d_{\mathrm{TV}}(P,Q) \approx \sup_{A \in \mathcal{A}} |P(A) - Q(A)|,$$

where $\mathcal{A}$ are sets of the form

$$\mathcal{A} = \{A(\mathbb{P}_1, \mathbb{P}_2) : \mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}\}, \tag{6}$$

and $A(\mathbb{P}_1, \mathbb{P}_2) = \{x : \mathbb{P}_1(x) > \mathbb{P}_2(x)\}$. This approximation allows us to construct statistically optimal estimators for $\mathcal{G}_{p,k}$.

3.2.1 Non-Asymptotic Robust Estimators for $\mathcal{G}_{p,k}$

Given samples $\{x^{(i)}\}_{i=1}^n$ from the mixture model $\mathbb{P}_\epsilon$ defined in (1), define $\widehat{\mathbb{P}}_{n,\epsilon}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{x^{(i)} \in A\right\}$ for all $A \in \mathcal{A}$, where $\mathcal{A}$ is the same as defined in (6) with the class of distributions $\mathcal{G}_{p,k}$. Our estimator is defined as

$$\mathbb{P}_{\widehat{\theta}} = \operatorname*{argmin}_{\mathbb{P}_\theta \in \mathcal{G}_{p,k}} \sup_{A \in \mathcal{A}} \left|\mathbb{P}_\theta(A) - \widehat{\mathbb{P}}_{n,\epsilon}(A)\right|. \tag{7}$$

The following lemma characterizes the performance of our estimator.

Lemma 2. *Given n samples from a contaminated distribution P_ϵ, the Yatracos estimate (7) satisfies with probability least $1-\delta$:*

$$d_{\mathrm{TV}}(\mathbb{P}_{\widehat{\theta}}, \mathbb{P}_{\theta^\star}) \leq 2\epsilon + \mathcal{O}\left(\sqrt{\frac{k\log(p^2e/k)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

The lemma above shows that the Yatracos estimate is close to the true Ising model in TV distance with high-probability. Combining Lemma 2 and Theorem 1, we get parameter error guarantees for the Yatracos estimate.

Corollary 2. *Given n samples from $\mathbb{P}_\epsilon$, the Yatracos' estimator returns a $\widehat{\theta}$ such that with probability at least $1 - \delta$,*

$$\|\widehat{\theta}(i) - \theta^\star(i)\|_2 \lesssim 2\epsilon\sqrt{\log(1/\epsilon)} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{k\log(p^2e/k)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) \quad \text{for all } i \in [p], \tag{8}$$

where $\tilde{\mathcal{O}}(.)$ hides logarithmic factors involving its argument.

Remarks. Note that the proposed estimator achieves the same (asymptotic) dimension-independent error as the Minimum Distance Estimate discussed in Section 3.1, while simultaneously achieving an $\widetilde{\mathcal{O}}\left(\sqrt{\frac{k\log p}{n}}\right)$ error rate. Moreover, observe that in the uncontaminated setting, i.e., when $\epsilon = 0$, this is the *first* estimator to get an $\widetilde{O}\left(\sqrt{\frac{k\log p}{n}}\right)$ error rate. As a consequence, Yatracos' estimator followed by an additional thresholding step gives the first estimator to recover the true edge set $E(\theta^\star)$ with only $\widetilde{\mathcal{O}}\left(\frac{k\log(p)}{\lambda^2}\right)$ samples. In contrast, the estimator proposed by [45] posit that the sample size should satisfy $\mathcal{O}(1/\lambda^4)$ when the parameters are unknown. In the contaminated case, note that we show a better dependence on ϵ – $\mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)})$ vs. $\sqrt{\epsilon}$ in [35]. The proof for Lemma 2 is presented in Section D.2 of the appendix. A similar analysis was conducted in [14], however [14] study density estimation, and not parameter estimation. The bound on the modulus of continuity obtained in Theorem 1 allows us to relate the TV distance between the estimated distribution and the true distribution to the parameter error, thus giving us bounds for parameter estimation.

3.2.2 Non-Asymptotic Robust Estimators for $\mathcal{G}_{p,d}$

Under the same setting as considered for $\mathcal{G}_{p,k}$, we see that directly employing the estimator (7) would lead to a sub-optimal rate. Our guarantee for (7) for $\mathcal{G}_{p,k}$ relies on the fact that parameters for Ising models in $\mathcal{G}_{p,k}$ contain at most k non-zero elements, hence the subset $A(\theta^{(1)}, \theta^{(2)}) = \{x : \mathbb{P}_{\theta^{(1)}}(x) > \mathbb{P}_{\theta^{(2)}}(x)\}$ is a half-space defined by a vector with at most $2k + 1$ non-zero elements. However, these subsets defined with parameters $\theta^{(1)}, \theta^{(2)}$ of two Ising models in $\mathcal{G}_{p,d}$ is a half-space defined by a vector that have at most $pd + 1$ non-zero elements. This leads to a rate term that is proportional to $\sqrt{pd\log(p)/n}$, which does not scale well in high-dimensional settings.

4 Robust Conditional Likelihood Estimators

In the previous section, we have seen that the estimator based on Yatracos classes [50] provides an approximate TV projection for $\mathcal{G}_{p,k}$ but not for $\mathcal{G}_{p,d}$. The main caveat with this estimator is that it is not tractable and takes infinite time. To circumvent this issue, we consider a more direct approach to robust estimation: we "robustify" the gradient samples obtained from samples $\{x^{(i)}\}_{i=1}^n$ of the contaminated distribution $\mathbb{P}_\epsilon = (1 - \epsilon)\mathbb{P}_{\theta^\star} + \epsilon Q$.

Neighborhood-based logistic regression. In a classical paper, Besag [4] made the key structural observation that under model (2), the conditional distribution of node X_i given the other variables $X_{-i} = x_{-i}$ is given by

$$\mathbb{P}_{\theta^\star}(X_i = x_i | X_{-i} = x_{-i}) = \frac{\exp(2x_i \sum_{t\in\mathcal{N}(i)} \theta^\star_{it} x_t)}{\exp(2x_i \sum_{t\in\mathcal{N}(i)} \theta^\star_{it} x_t) + 1} = \sigma(x_i \langle 2\theta^\star(i), x_{-i}\rangle).$$

Thus the variable X_i can be viewed as the response variable in a logistic regression model with X_{-i} as the covariates and $2\theta^\star(i)$ as the regression vector. In particular, this implies that $\mathbb{E}_{x\sim\mathbb{P}_{\theta^\star}}[\nabla l_i(2\theta^\star(i); x)] = \mathbf{0}$ where $\ell_i(\theta(i); x) = \log\sigma(x_i \langle\theta(i), x_{-i}\rangle)$ is the conditional log-likelihood of x under $\mathbb{P}_\theta$. Note that for graphs with maximum degree at most d, the parameter vector $\theta^\star(i)$ has at most d non-zero entries, and for graphs with at most k edges, the parameter vector $\theta^\star(i)$ has at most k non-zero entries. Ravikumar et al. [42] solved an ℓ_1-regularized logistic regression to recover the node parameters for graphs with bounded maximum degree. However, in our setting, the data is contaminated with outliers, and hence the minimizer of the likelihood can be arbitrarily bad. While there has been recent work giving provably optimal algorithms for robust logistic regression [41], all of these results are in the low-dimensional setting. We propose the *first* statistically optimal estimator for sparse logistic regression, and use that to provide estimators for learning Ising models.

Algorithm 1 Robust1DMean - Robust univariate mean estimator

```
Require: Samples $\{z^{(i)}\}_{i=1}^{2n}$, corruption level $\epsilon$, confidence level $\delta$
Split $\{z^{(i)}\}_{i=1}^{2n}$ into two subsets $\mathcal{Z}_1 = \{z^{(i)}\}_{i=1}^{n}$ and $\mathcal{Z}_2 = \{z^{(i)}\}_{i=n+1}^{2n}$
Set $\beta = \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right)$
$n_1 = n\left(1 - 2\beta - \sqrt{2\beta\log(4/\delta)/n} - \log(4/\delta)/n\right)$
Using $\mathcal{Z}_1$, identify $\widehat{I} = [a, b]$ which is the shortest interval containing $n_1$ points
return $\frac{1}{n_2}\sum_{i=n+1}^{2n} z^{(i)}\mathbb{I}\left\{z^{(i)} \in \widehat{I}\right\}$ where $n_2 = \sum_{i=n+1}^{2n}\mathbb{I}\left\{z^{(i)} \in \widehat{I}\right\}$
```

Robust Sparse Logistic Regression. Our approach is based on a reduction to robust univariate estimation initially proposed by [40]. In particular, note that when we have clean data, then, in the population setting, $\theta^\star(i)$ is the unique solution to the equation $\|\mathbb{E}_{x\sim\mathbb{P}_{\theta^\star}}[\nabla\ell_i(\theta(i);x)]\|_2 = \mathbf{0}$ or equivalently, it is the unique minimizer for the following optimization problem:

$$\theta^\star(i) = \underset{w:\|w\|_0\le s}{\operatorname{argmin}} \sup_{u\in\mathcal{S}^{p-2}} \left|\mathbb{E}_{x\sim\mathbb{P}_{\theta^\star}}[u^T\nabla\ell_i(w;x)]\right|,$$

where we have simply used the variational form of the norm of a vector. Observe that $\mathbb{E}_{x\sim\mathbb{P}_{\theta^\star}}[u^T\nabla\ell_i(w;x)]$ is simply the population (uncontaminated) mean of the gradients, when projected along the direction u. Unfortunately, we only have finite samples which are moreover contaminated. We can pass these univariate projections of the gradient through a *robust* univariate mean estimator, and return a point which has the *smallest* (robust) mean along any direction. This leads to the following program,

$$\widehat{\theta}(i) = \underset{w\in\mathcal{N}_s^\gamma(\mathcal{S}^{p-2})}{\operatorname{argmin}} \sup_{u\in\mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-2})} \left|\mathsf{Robust1DMean}(\{u^T\nabla\ell_i(w;x^{(j)})\}_{j=1}^n)\right|, \tag{9}$$

where $\mathcal{N}_s^\gamma(\mathcal{S}^{p-2})$ is a γ-cover of the unit sphere over $p-1$ dimensions with s non-zero entries i.e., for every $x \in \mathcal{S}^{p-2}$ that has s non-zero entries, there exists $y \in \mathcal{N}_s^\gamma(\mathcal{S}^{p-2})$ such that $\|x-y\|_2 \le \gamma$. Our robust univariate mean estimator is based on the shortest interval estimator (Shorth) studied in [2, 34, 40]. The estimator, presented in Algorithm 1, proceeds by using half of the samples to identify the shortest interval containing roughly $(1-\epsilon)n$ fraction of the points, and then the remaining half of the points is used to return an estimate of the mean. Intuitively, this estimator effectively trims distant outliers, thereby limiting their influence on the estimate.

We assume that the contamination level ϵ, confidence parameter δ, and sparsity s are such that,

$$2\epsilon + \sqrt{\epsilon\left(\frac{s\log(p)}{n} + \frac{\log(p/\delta)}{n}\right)} + \frac{s\log(p)}{n} + \frac{\log(4p/\delta)}{n} < c, \tag{10}$$

for some small constant $c > 0$. As noted earlier, the sparsity parameter s is the maximum degree d for $\mathcal{G}_{p,d}$ and the maximum number of edges k for $\mathcal{G}_{p,k}$.

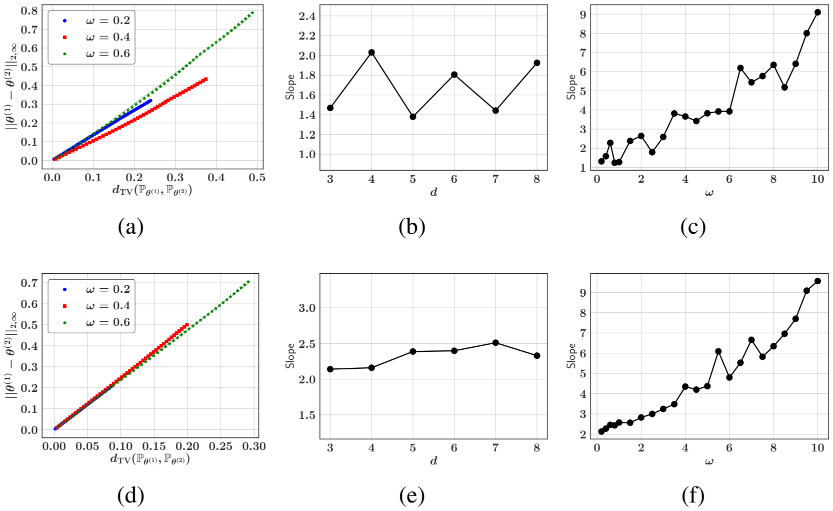
Theorem 2 (Guarantees for $\mathcal{G}_{p,d}$)**.** *Under the setting considered in 4 along with Assumption (3), the estimator in (9) returns estimates* $\{\widehat{\theta}(i)\}_{i=1}^p$ *with* $\gamma = \max\left\{\frac{\epsilon}{p}, \frac{\log(1/\delta)}{np}\right\}$ *returns with probability at least* $1-\delta$

$$\|\widehat{\theta}(i)-\theta^\star(i)\|_2 \lesssim \epsilon\sqrt{C(\alpha)\log\left(\frac{1}{\epsilon}\right)} + \sqrt{C(\alpha)\frac{d}{n}\log\left(\frac{3ep^2}{d\gamma}\right)} + \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right) \text{ for all } i \in [p].$$

Corollary 3 (Guarantees for $\mathcal{G}_{p,k}$)**.** *Under the setup considered in Theorem 2, the estimator in (9) returns estimates* $\{\widehat{\theta}(i)\}_{i=1}^p$ *with* $\gamma = \max\left\{\frac{\epsilon}{p}, \frac{\log(1/\delta)}{np}\right\}$ *returns with probability at least* $1-\delta$

$$\|\widehat{\theta}(i)-\theta^\star(i)\|_2 \lesssim \epsilon\sqrt{C(\alpha)\log\left(\frac{1}{\epsilon}\right)} + \sqrt{C(\alpha)\frac{k}{n}\log\left(\frac{3ep^2}{k\gamma}\right)} + \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right) \text{ for all } i \in [p].$$

Remarks. Observe that our estimator achieves the same (asymptotic) bias as the Minimum Distance Estimator, previously discussed in Section 3.1. Define the recovered edge set as those edges (i, j)



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1: Left: Variation of $\|\theta^{(1)} - \theta^{(2)}\|_{2,\infty}$ with $d_{\text{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}})$ for $G^{(1)}, G^{(2)} \in \mathcal{G}^{\text{clique}}_{15,4}$ (top) and $G^{(1)}, G^{(2)} \in \mathcal{G}^{\text{star}}_{15,4}$ (bottom) graphs with varying ω. Middle: Variation of slope with d for cliques (top) and star (bottom) with $p = 12$ and $\omega = 0.4$. Right: Variation of slope with ω for cliques (top) and star (bottom) with $p = 15$ and $d = 5$. The slope is defined as $\frac{\|\theta^{(1)} - \theta^{(2)}\|_{2,\infty}}{d_{\text{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}})}$.

satisfying $|\widehat{\theta}_{ij}| \geq \lambda/2$. When $\epsilon = 0$, i.e., no contamination, for $\mathcal{G}_{p,d}$, we require the number of samples $n \geq \mathcal{O}\left(\frac{d \log(p)}{\lambda^2}\right)$ to recover the true edge set $E(\theta^\star)$. Even in the uncontaminated setting, there is *no* known estimator which achieves the same optimal sample complexity as ours. In particular, Santhanam and Wainwright [45] achieve similar rates when they assume that the structure is known, while other approaches of [36, 42] have worse dependence on the degree d. Hence, our proposed estimator has an optimal (asymptotic) bias and optimal high probability bounds. For $\mathcal{G}_{p,k}$, we obtain the same rate and sample complexity as Yatracos' estimator (7), which we remarked is optimal. The proof of Theorem 2 is presented in Section E.1 of the appendix.

5 Synthetic Experiments

Our theoretical results crucially hinge on bounds on the TV modulus of continuity derived in Theorem 1, and we devote this section to corroborating these bounds.

Setup. We consider two different ensembles. A graph $G \in \mathcal{G}^{\text{star}}_{p,d}$ when one of the p nodes is connected d other vertices, and no other edges are present in the graph, resembling a star. A graph $G \in \mathcal{G}^{\text{clique}}_{p,d}$ contains $\lfloor \frac{p}{d+1} \rfloor$ cliques of size $d+1$, and the remainder of the nodes $p \mod (d+1)$ fully connected amongst themselves. We generate our plots in the following manner: first we construct two graphs with the same structure - either from $\mathcal{G}^{\text{clique}}_{p,d}$ of $\mathcal{G}^{\text{star}}_{p,d}$. We instantiate parameters for the first graph with $\theta^{(1)}$ with model width ω and then vary the parameters for the second graph as $\theta^{(2)} = \theta^{(1)} \cdot \frac{i}{25}$ for i ranging from 1 to 50. We vary $p \in \{12, 15\}$, $d \in \{3 : 8 : 1\}$ and $\omega \in \{0.2 : 1.0 : 0.2\} \cup \{1.5 : 10 : 0.5\}$ where $\{a : b : c\}$ denotes values between a and b (both inclusive) with consecutive values differing by c.

Results. Figures 1(a) and 1(d) exhibits a linear relationship between $d_{\text{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}})$ and $\|\theta^{(1)} - \theta^{(2)}\|_{2,\infty}$, as suggested by our theoretical results from previous sections. Furthermore, we notice that the slope is not drastically affected by ω, which also suggests that the constant $C(\alpha)$ appearing in our results is $O(1)$. We also note from Figures 1(b) and 1(e), that the slope is unaffected by a change in degree. Finally, in Figures 1(c) and 1(f), we notice the variation in the slope with increasing model width ω. While our current result study the case when $\omega < 1$, it is also interesting to note an increasing trend when $\omega \geq 1$ suggesting an explicit dependence on ω in the low-temperature regime.

6 Discussion and Future Work

In this work we provided the first statistically optimal robust estimators for learning Ising models in the high temperature regime. Our estimators achieved optimal asymptotic error in the ϵ-contamination model, and also high-probability deviation bounds in the uncontaminated setting. There are several avenues for future work, some of which we discuss below.

Beyond Dobrushin's conditions. In the low-temperature setting, Lindgren et al. [35] showed the existence of an estimator which gets an $O(\sqrt{\epsilon})$ error. In Appendix A, we tighten this for edge-bounded graphs by providing estimators which achieve $O(\min(\sqrt{\epsilon}, \epsilon\sqrt{k}))$ error, where k is the maximum number of edges in the graph. However, giving matching lower bounds in this setting is an open problem. Our synthetic experiments surprisingly show that one may expect similar rates in the two temperature regimes.

Computationally Efficient Estimators. While in this work, we designed statistically optimal estimators that achieve an $O(\epsilon\sqrt{\log(1/\epsilon)})$ parameter error, whereas, existing computationally efficient approaches [30, 35] achieve a sub-optimal error of $O(\sqrt{\epsilon})$. Developing computationally efficient algorithms which close this gap is an interesting open problem.

Other Contamination Models. In this work, our focus was on designing estimators for the ϵ-contaminated model, i.e., where a fraction of the data is arbitrarily corrupted. Another model of corruption - motivated by sensor networks and distributed computation where node failures are common - is when only a few features(nodes) get corrupted, and we still want to learn the appropriate graph structure for the uncontaminated nodes.

Broader Impact

In this work, we provide statistically optimal estimators for learning Ising models under contamination. Ising models are themselves used in a variety of domains to learning relationship between pairs of binary random variables. One extremely interesting application is in the field of opinion analysis and voting network analysis. For instance, the nodes of the graph represent the voting base and the samples given to us are votes made of a series of topics as obtained via polls. Such estimators will help capture associations between voters. However, in a day and age where voting patterns are susceptible to adversarial corruptions, it is safe to assume that the vote samples are corrupted too. Using standard methods such as ℓ_1-regularized logistic regression could have the unintended consequence of amplifying the biases from corrupted data, leading to poor judgements, whereas our methods are optimal resilient to such corruptions. However, if used without prior analysis of the data presented, this could potentially reduce the effect of outlier samples, which in the case of voting patterns, are representative of a minority groups.

Acknowledgements

AP, VS and PR acknowledge the support of NSF via IIS-1955532, OAC-1934584, DARPA via HR00112020006, and ONR via N000141812861. SB and AP acknowledge the support of NSF via DMS-17130003 and CCF-1763734. We would also like to thank an anonymous reviewer for pointing out related work by Devroye et al. [14].

References

[1] Mehmet Eren Ahsen and Mathukumalli Vidyasagar. An approach to one-bit compressed sensing based on probably approximately correct learning theory. *The Journal of Machine Learning Research*, 20(1):408–430, 2019.

[2] DF Andrews, PJ Bickel, FR Hampel, PJ Huber, WH Rogers, and JW Tukey. Robust estimates of location: Survey and advances, 1972.

[3] Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212, 2017.

[4] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

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

[6] 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.

[7] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017.

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

[9] George R Cross and Anil K Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):25–39, 1983.

[10] Imre Csiszár and Zsolt Talata. Consistent estimation of the basic neighborhood of markov random fields. *The Annals of Statistics*, pages 123–145, 2006.

[11] Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Anthimos Vardis Kandiros. Estimating ising models from one sample. *arXiv preprint arXiv:2004.09370*, 2020.

[12] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. *IEEE Transactions on Information Theory*, 65(11):6829–6852, 2019.

[13] Christopher De Sa, Kunle Olukotun, and Christopher Ré. Ensuring rapid mixing and low bias for asynchronous gibbs sampling. In *JMLR workshop and conference proceedings*, volume 48, page 1567. NIH Public Access, 2016.

[14] Luc Devroye, Abbas Mehrabian, Tommy Reddad, et al. The minimax learning rates of normal and ising undirected graphical models. *Electronic Journal of Statistics*, 14(1):2338–2361, 2020.

[15] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

[16] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceeds of the 34th International Conference on Machine Learning*, pages 999–1008, 2017.

[17] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[18] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1596–1606, 2019.

[19] PL Dobruschin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability & Its Applications*, 13(2):197–224, 1968.

[20] Roland L Dobrushin and Senya B Shlosman. Completely analytical interactions: constructive description. *Journal of Statistical Physics*, 46(5-6):983–1014, 1987.

[21] David L Donoho and Richard C Liu. The" automatic" robustness of minimum distance functionals. *The Annals of Statistics*, pages 552–586, 1988.

[22] David L Donoho and Richard C Liu. Geometrizing rates of convergence, iii. *The Annals of Statistics*, pages 668–701, 1991.

[23] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303 (5659):799–805, 2004.

[24] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741, 1984.

[25] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Higher order concentration for functions of weakly dependent random variables. *Electron. J. Probab.*, 24:19 pp., 2019.

[26] Martin Hassner and Jack Sklansky. The use of markov random fields as models of texture. In *Image Modeling*, pages 185–198. Elsevier, 1981.

[27] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[28] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

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

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

[31] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046. ACM, 2018.

[32] Christof Külske. Concentration inequalities for functions of gibbs fields with application to diffraction and random gibbs measures. *Communications in mathematical physics*, 239(1-2): 29–51, 2003.

[33] H Künsch. Decay of correlations under dobrushin's uniqueness condition and its applications. *Communications in Mathematical Physics*, 84(2):207–222, 1982.

[34] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 665–674. IEEE, 2016.

[35] Erik M Lindgren, Vatsal Shah, Yanyao Shen, Alexandros G Dimakis, and Adam Klivans. On robust learning of ising models. *NeurIPS Workshop on Relational Representation Learning*, 2019.

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

[37] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. 1999.

[38] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

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

[40] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019.

[41] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, Pradeep Ravikumar, et al. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B*, 82 (3):601–627, 2020.

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

[43] Brian D Ripley. *Spatial statistics*, volume 575. John Wiley & Sons, 2005.

[44] Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[45] 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.

[46] Daniel W Stroock and Boguslaw Zegarlinski. The logarithmic sobolev inequality for discrete spin systems on a lattice. *Communications in Mathematical Physics*, 149(1):175–193, 1992.

[47] 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.

[48] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[49] Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. In *Advances in Neural Information Processing Systems*, pages 8069–8079, 2019.

[50] Yannis G Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, pages 768–774, 1985.

A Beyond Dobrushin's Conditions.

All of our previous results are under the high temperature condition (3), where we rely of special properties of Ising models namely sub-Gaussianity of Ising models random variables. Following this effort, we attempt to analyze classes of Ising models where this condition doesn't hold to present an even more general analysis. Towards this end, we present moduli of continuity bounds as presented in Theorem 1. Here, we look out for dependence in the model width parameter in addition to the effective dimensionality of the problem (d in the case of $\mathcal{G}_{p,d}$ and k in the case of $\mathcal{G}_{p,k}$, and the tolerance parameter ϵ.

Theorem 3. *Consider two Ising models defined over two graphs $G^{(1)}$ and $G^{(2)}$ over p vertices with parameters $\theta^{(1)}$ and $\theta^{(2)}$ respectively, satisfying $\omega(\theta^{(1)}), \omega(\theta^{(2)}) \leq \omega$. If $d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}) \leq \epsilon$, then we have the following results for all $i \in [p]$:*

(a) If $G^{(1)}, G^{(2)} \in \mathcal{G}_{p,d}$, then

$$\|\theta^{(2)}(i) - \theta^{(1)}(i)\|_2 \lesssim \min\{\sqrt{\epsilon}, \epsilon\sqrt{d}\}\, \omega \exp(O(\omega)). \tag{11a}$$

(b) If $G_1, G_2 \in \mathcal{G}_{p,k}$, then

$$\|\theta^{(2)}(i) - \theta^{(1)}(i)\|_2 \lesssim \min\{\sqrt{\epsilon}, \epsilon\sqrt{k}\}\, \omega \exp(O(\omega)). \tag{11b}$$

Similar to Theorem 3, we get a modulus of continuity bound for the loss function defined by the $(2, \infty)$-norm. Note that as ϵ tends to 0, the bounds also tend to 0. However, it is worth noting that our primitive analysis contains an additional factor in d/k based on the graph class considered. The sub-optimality is clear when we set $\omega = O(1)$, and the bounds while retaining a optimal dependence on ϵ have an additional dependence with d/k when compared to the result in Theorem 1. Our analysis of the Yatracos estimator (7) does not depend of any specific bounds on the model width, and hence with the derived modulus of continuity bound, we arrive at the following corollary for the estimation error of the Yatracos estimate:

Corollary 4. *Given n samples from the distribution $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}_{\theta^\star} + \epsilon Q$, where $\mathbb{P}_{\theta^\star} \in \mathcal{G}_{p,k}(\lambda, \omega)$ and Q is an arbitrary distribution supported over $\{-1, +1\}^p$, the parameter of Yatracos estimate (7) satisfies:*

$$\|\widehat{\theta}(i) - \theta^\star(i)\|_2 \lesssim \sqrt{k}\omega e^{\mathcal{O}(\omega)}\epsilon + \mathcal{O}\left(k\omega e^{\mathcal{O}(\omega)}\sqrt{\frac{\log(p^2 e/k)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) \quad \text{for all } i \in [p].$$

Note that as $n \to \infty$, the bias of the estimator has optimal dependence on ϵ, but incurs an additional dependence of $\sqrt{k}$. For $\epsilon = 0$ *i.e.* no contamination, the rate we achieve is approximately $\omega e^{\mathcal{O}(\omega)} k \sqrt{\frac{\log(p)}{n}}$, which leads to the number of samples $n \geq \mathcal{O}\left(\frac{k^2\omega^2 e^{\mathcal{O}(\omega)}\log(p)}{\lambda^2}\right)$ required to recover the true edge set $E(\theta^\star)$, and this is comparable to existing sample complexity results for learning Ising models belonging to $\mathcal{G}_{p,k}(\lambda, \omega)$ [45]. We present the proof of Theorem 3 in Section F.

B Useful Properties of Ising models

In this section, we summarize some useful properties of Ising models which we use judiciously in our proofs. These results have appeared in previous work, but we state them for the sake of completeness.

B.1 Sub-Gaussianity of Ising model distributions in the high temperature regime

First, we present a result from [25], which states that a random variable distributed according to an Ising model in the high temperature regime is sub-Gaussian.

Proposition 3 ([25, Theorem 1.4])**.** *Let $z \sim \mathbb{P}$ be a random variable whose distribution $\mathbb{P}$ is an Ising model over p nodes in the high temperature regime (3) with constant α. Then for $v \sim \mathbb{R}^p$:*

$$\Pr_{z\sim\mathbb{P}}\left(|\langle v, z\rangle| > t\right) \leq 2\exp\left(-\frac{t^2}{C(\alpha)||v||_2^2}\right), \tag{12}$$

where $C(\alpha)$ is a constant depending on α.

B.2 Strong convexity of the negative conditional log-likelihood

Here we present a proposition that states that the population negative conditional log-likelihood is strongly convex. This proposition is obtained using a result by Dagan et al. [11]. We first state the result by Dagan et al. [11] below, and then use it to show that the population negative condition log-likelihood is strongly convex.

Proposition 4 ([11, Lemma 10])**.** *Let z be a random variable distributed w.r.t. an Ising model over p nodes whose parameter θ satisfies $\max_{i\in[p]} \|\theta(i)\|_\infty \leq \omega$ and $\min_{i\in[p]} \mathbb{P}_\theta(X_i = 1|X_{-i} = x_{-i})(1 - \mathbb{P}_\theta(X_i = 1|X_{-i} = x_{-i})) \geq \gamma$. Then for any $v \in \mathbb{R}^p$, we have that:*

$$\mathrm{Var}[\langle v, z\rangle] \geq \frac{C_1\gamma^2||v||_2^2}{\omega},$$

where C_1 is a universal constant.

Now, let $\mathcal{L}_{\theta,i}(w)$ be the population negative conditional log-likelihood for node X_i, where X is sampled from the Ising model distribution $\mathbb{P}_\theta$. Formally, $\mathcal{L}_{\theta,i}(w) = -\mathbb{E}_{z\sim\mathbb{P}_\theta}[\ell_i(w; z)]$, where $\ell_i(w; z)$ is the conditional log-likelihood of z under $\mathbb{P}_\theta$ with respect to the i^{th} node. As stated earlier, by the maximum likelihood principle, $\nabla\mathcal{L}_{\theta,i}(2\theta(i)) = \mathbf{0}$. With this definition, we have the Hessian of the population negative conditional log-likelihood as $\nabla^2\mathcal{L}_{\theta,i}(w) = \mathbb{E}_{z\sim\mathbb{P}_\theta}[\nabla^2\ell_i(w; z)]$. Then, we have the following result.

Proposition 5. *Let $\mathbb{P}_\theta$ be an Ising model over p nodes whose parameter satisfies $\max_{i\in[p]} \|\theta(i)\|_\infty \leq \omega$, and let $w \in \mathbb{R}^{p-1}$ be such that $\|w\|_1 \leq 2\omega$. Then, for any vector $v \in \mathbb{R}^{p-1}$, there exists a universal constant $C > 0$ such that:*

$$v^T\nabla^2\mathcal{L}_{\theta,i}(w)v \geq C\frac{\exp(-O(\omega))}{\omega}\|v\|_2^2.$$

Proof. First, observe that

$$\begin{aligned}\nabla^2\mathcal{L}_{\theta,i}(w) &= \mathbb{E}_{z\sim\mathbb{P}_\theta}\left[\sigma(z_i\langle w, z_{-i}\rangle)(1-\sigma(z_i\langle w, z_{-i}\rangle))z_{-i}z_{-i}^T\right] \\ \Rightarrow v^T\nabla^2\mathcal{L}_{\theta,i}(w)v &= \mathbb{E}_{z\sim\mathbb{P}_\theta}\left[\sigma(z_i\langle w, z_{-i}\rangle)(1-\sigma(z_i\langle w, z_{-i}\rangle))\langle z_{-i}, v\rangle^2\right].\end{aligned}$$

In Lemma 6, we show that for any $\|w\|_1 \leq 2\omega$, we have that

$$\sigma(z_i\langle w, z_{-i}\rangle)(1-\sigma(z_i\langle w, z_{-i}\rangle)) \geq \frac{\exp(-2\omega)}{4}. \tag{13}$$

We now lower bound $\mathbb{E}[\langle z_{-i}, v\rangle^2]$. Since Ising model has zero mean field, we have that $\mathbb{E}[\langle z_{-i}, v\rangle^2] = \mathrm{Var}[\langle z_{-i}, v\rangle]$. Furthermore, due the assumptions placed on the parameter of the Ising model, we obtain that for any $x \in \{-1,+1\}^{p-1}$, $\mathbb{P}_\theta(X_i = 1|X_{-i} = x)(1 - \mathbb{P}_\theta(X_i = 1|X_{-i} = x)) \geq \frac{1}{4}\exp(-4\omega)$. This can be shown as follows. For any $z \in \{-1,+1\}$ and $x \in \{-1,+1\}^{p-1}$, we have that:

$$\begin{aligned}
\mathbb{P}_\theta(X_i = z|X_{-i} = x) &= \frac{1}{1+\exp(-z\left\langle 2\theta(i,-i), x\right\rangle)} \\
&\overset{(i)}{\geq} \frac{1}{1+\exp(2\omega)} \\
&\geq \frac{1}{2\exp(2\omega)} = \frac{\exp(-2\omega)}{2} \\
\Rightarrow \mathbb{P}_\theta(X_i = 1|X_{-i} = x)\mathbb{P}_\theta(X_i = 0|X_{-i} = x) &\geq \frac{\exp(-2\omega)}{2}\frac{\exp(-2\omega)}{2} \\
&= \frac{\exp(-4\omega)}{4}
\end{aligned}$$

where Step (i) uses Hölder's inequality as: $|\left\langle 2\theta(i,-i), x\right\rangle| \leq 2\omega \Rightarrow -z\left\langle 2\theta(i,-i), x\right\rangle \leq 2\omega$.

Using this in Proposition 4, we have that:

$$\mathrm{Var}[\langle v, z_{-i}\rangle] \geq C\frac{\exp(-8\omega)\|v\|_2^2}{\omega} \tag{14}$$

where C is a universal constant.

Combining (13) and (14), we obtain the statement of the lemma. □

B.2.1 Auxiliary Lemmata

Lemma 6. *If $w \in \mathbb{R}^{p-1}$ such that $||w||_1 \leq 2\omega$, then for $x, y \in \{-1,+1\}^{p-1} \times \{-1,+1\}$:*

$$\sigma(y\langle w,x\rangle)(1-\sigma(y\langle w,x\rangle)) = \frac{\exp(-y\langle w,x\rangle)}{(1+\exp(-y\langle w,x\rangle))^2} \geq \frac{\exp(-|y\langle w,x\rangle|)}{4} \geq \frac{\exp(-2\omega)}{4} \tag{15}$$

Proof. Consider $f(a) = \sigma(a)(1-\sigma(a)) = \frac{\exp(-a)}{(1+\exp(-a))^2} = \frac{\exp(a)}{(1+\exp(a))^2}$. Now for $a > 0$:

$$e^{-a} < 1 \Leftrightarrow e^{-a} + 1 < 2 \Leftrightarrow (e^{-a}+1)^2 < 4 \Leftrightarrow \frac{\exp(-a)}{(1+\exp(-a))^2} \geq \frac{\exp(-a)}{4}$$

For $a < 0$:

$$e^{a} < 1 \Leftrightarrow e^{a} + 1 < 2 \Leftrightarrow (e^{a}+1)^2 < 4 \Leftrightarrow \frac{\exp(a)}{(1+\exp(a))^2} \geq \frac{\exp(a)}{4}$$

Therefore:

$$f(a) \geq \frac{\exp(-|a|)}{4}$$

By Hölder's inequality, $|y\langle w,x\rangle| \leq ||w||_1||x||_\infty \leq 2\omega$. This implies that

$$\sigma(y\langle w,x\rangle)(1-\sigma(y\langle w,x\rangle)) = f(y\langle w,x\rangle) \geq \frac{\exp(-|y\langle w,x\rangle|)}{4} \geq \frac{\exp(-2\omega)}{4}$$

□

C Proofs of Propositions in Section 2

In this section, we present the proofs for Theorem 1 and Lemma 1.

C.1 Proof of Theorem 1

Here, we derive bounds on the modulus of continuity defined in (4) with the loss function given by the $\ell_{2,\infty}$ norm of the parameters.

Proof Sketch. We begin by giving a brief proof outline. $\mathbb{P}_{\theta^{(1)}}$ and $\mathbb{P}_{\theta^{(2)}}$ are two Ising models in the high temperature regime (3) with constant α, and additionally satisfy $d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}) \leq \epsilon$. Consider $\mathcal{L}_{\theta^{(1)},i}$ to be the population negative conditional log-likelihood for the i^{th} node with respect to $\mathbb{P}_{\theta^{(1)}}$ defined earlier. We earlier noted that $\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(1)}(i)) = 0$ by the maximum likelihood principle.

In Lemma 7, we show that under these conditions, the gradient $\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))$ satisfies $\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\|_2 \leq \sqrt{C(\alpha)}\epsilon\sqrt{\log(1/\epsilon)}$, where $C(\alpha)$ is a universal constant only depending on α. With this intermediate result, we complete the proof of the theorem as follows. Considering the Taylor series expansion of $\mathcal{L}_{\theta^{(1)},i}$ around $2\theta^{(2)}(i)$, we get

$$\begin{aligned}
\mathcal{L}_{\theta^{(1)},i}(2\theta^{(1)}(i)) &= \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) + \left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \varDelta_i \right\rangle + \frac{1}{2}\varDelta_i^T \nabla^2 \mathcal{L}_{\theta^{(1)},i}(\widetilde{w})\varDelta_i \\
&\overset{(i)}{\geq} \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) + \left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \varDelta_i \right\rangle + \frac{C}{2}\frac{\exp(-O(\omega))}{\omega}\|\varDelta_i\|_2^2 \\
&\overset{(ii)}{\geq} \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) + \left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \varDelta_i \right\rangle + C'\frac{\exp(-c(1-\alpha))}{1-\alpha}\|\varDelta_i\|_2^2,
\end{aligned}$$

where $\widetilde{w}$ lies between $2\theta^{(2)}(i)$ and $2\theta^{(1)}(i)$, and $\varDelta_i = 2\theta^{(1)}(i) - 2\theta^{(2)}(i)$. In step (i), we have used the result in Proposition 5 and in step (ii) we use the fact that $\omega \leq 1 - \alpha$.

We also know by the maximum likelihood principle that $\mathcal{L}_{\theta^{(1)},i}(2\theta^{(1)}(i)) \leq \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))$, and substituting this in the inequality above yields

$$C'\frac{\exp(-c(1-\alpha))}{1-\alpha}\|\varDelta_i\|_2^2 \leq -\left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \varDelta_i \right\rangle \leq \left|\left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \varDelta_i \right\rangle\right|.$$

Finally, we bound the right hand side using the Cauchy-Schwarz inequality and the result from Lemma 7 to get

$$\left|\left\langle \nabla \mathcal{L}_{\theta^{(1)}(i)}(2\theta^{(2)}(i)), \varDelta_i \right\rangle\right| \leq \|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\|_2 \|\varDelta_i\|_2 \leq \sqrt{C(\alpha)}\epsilon\sqrt{\log(1/\epsilon)}\|\varDelta_i\|_2,$$

and substituting this in the quadratic bound above gives

$$\|\varDelta_i\|_2 \leq C_1(\alpha)\epsilon\sqrt{\log(1/\epsilon)}, \qquad C_1(\alpha) = \frac{1}{C'}(1-\alpha)\exp(c(1-\alpha))\sqrt{C(\alpha)}.$$

□

We now state Lemma 7 and prove it below.

Lemma 7. *Let $\mathbb{P}_{\theta^{(1)}}$ and $\mathbb{P}_{\theta^{(2)}}$ be two Ising models in the high temperature regime (3) with constant α that satisfies $d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}) \leq \epsilon$. Then, there exists a universal constant $C(\alpha)$ that only depends on α such that*

$$\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\|_2 \leq \sqrt{C(\alpha)}\epsilon\sqrt{\log(1/\epsilon)} \qquad \textit{for all } i \in [p]$$

Proof. Recall that $\mathcal{L}_{\theta^{(1)},i}(w) = \mathbb{E}_{z\sim\mathbb{P}_{\theta^{(1)}}}[\ell_i(w;z)]$. By the maximum likelihood principle, we know that

$$\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(1)}(i)) = \mathbf{0} \qquad \nabla \mathcal{L}_{\theta^{(2)},i}(2\theta^{(2)}(i)) = \mathbf{0}$$

Since $d_{\mathrm{TV}}(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}}) \leq \epsilon$, there exists an ϵ-coupling $\mathcal{C}$ between $\mathbb{P}_{\theta^{(1)}}$ and $\mathbb{P}_{\theta^{(2)}}$. In particular, $\mathcal{C}$ is a joint distribution over z_1, z_2 such that the respective marginals are $z_1 \sim \mathbb{P}_{\theta^{(1)}}$ and $z_2 \sim \mathbb{P}_{\theta^{(2)}}$, and $\mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\mathbb{I}\{z_1 \neq z_2\}] \leq \epsilon$.

The rest of the proof begins by making the observation that $\nabla\mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) = \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_1)]$. By introducing indicator random variables for the cases when z_1 and z_2 are equal or not, we have

$$\begin{aligned}\nabla\mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i)) &= \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_1)\mathbb{I}\{z_1 \neq z_2\}] + \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_1)\mathbb{I}\{z_1 = z_2\}] \\ &= \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_1)\mathbb{I}\{z_1 \neq z_2\}] + \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_2)\mathbb{I}\{z_1 = z_2\}] \\ &\overset{(a)}{=} \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_1)\mathbb{I}\{z_1 \neq z_2\}] - \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_2)\mathbb{I}\{z_1 \neq z_2\}],\end{aligned}$$

where step (a) follows from the stationarity of $2\theta^{(2)}(i)$ under $\mathbb{P}_{\theta^{(2)}}$ like so.

$$\begin{aligned}\mathbf{0} &= \nabla\mathcal{L}_{\theta^{(2)},i}(2\theta^{(2)}(i)) \\ &= \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_2)] \\ &= \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_2)\mathbb{I}\{z_1 = z_2\}] + \mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\nabla\ell_i(2\theta^{(2)}(i); z_2)\mathbb{I}\{z_1 \neq z_2\}].\end{aligned}$$

Therefore, for any vector $v \in \mathcal{S}^{p-2}$, we have that

$$\begin{aligned}\left|\left\langle v, \nabla\mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| &= \Big|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\left\langle v, \nabla\ell_i(2\theta^{(2)}(i); z_1)\right\rangle \mathbb{I}\{z_1 \neq z_2\}] \\ &\quad -\mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\left\langle v, \nabla\ell_i(2\theta^{(2)}(i); z_2)\right\rangle \mathbb{I}\{z_1 \neq z_2\}]\Big| \\ &\leq \underbrace{\left|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\left\langle v, \nabla\ell_i(2\theta^{(2)}(i); z_1)\right\rangle \mathbb{I}\{z_1 \neq z_2\}]\right|}_{T_1} \\ &\quad + \underbrace{\left|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}[\left\langle v, \nabla\ell_i(2\theta^{(2)}(i); z_2)\right\rangle \mathbb{I}\{z_1 \neq z_2\}]\right|}_{T_2}.\end{aligned}$$

Bounding T_2: Note that $\nabla\ell_i(w; z_1) = (\sigma(\langle w, z_1(-i)\rangle z_1(i)) - 1)z_1(-i)z_1(i)$. Since $z_1 \sim \{-1,+1\}^p$, we have that $|(\sigma(\langle w, z_1(-i)\rangle z_1(i)) - 1)z_1(i)| < 1$, and hence we get $|\langle v, \nabla\ell_i(w; z_1)| < |\langle v, z_1(-i)|$.

This in turn implies

$$\Pr(|\langle v, \nabla\ell_i(w; z_1)\rangle| > t) \leq \Pr(|\langle v, z_1(-i)\rangle| > t) \overset{(b)}{\leq} 2\exp\left(-\frac{t^2}{C(\alpha)}\right)$$

where step (b) follows from the sub-Gaussianity of random variables distributed with respect to an Ising model in the high temperature regime (Proposition 3). Using standard tail bounds (see [48, Chapter 2]), we obtain that $\mathbb{E}[\exp(\lambda(\langle v, \nabla\ell_i(w; z_1)\rangle))] \leq \exp\left(\frac{C\lambda^2 C(\alpha)}{2}\right)$. To finally bound T_2, we use the following result from [38].

Proposition 8 ([38, Lemma 2.3]). *Let Z be a random variable such that $\mathbb{E}[\exp(\lambda Z)] \leq e^{\frac{\lambda^2\sigma^2}{2}}$. For any measurable event A, we have*

$$|\mathbb{E}[Z \cdot \mathbb{I}\{A\}]| \leq \sigma P(A)\sqrt{\log(1/P(A))}.$$

In T_2, the event A is $z_1 \neq z_2$ and this occurs with probability less than ϵ. Hence, we get $T_2 \leq C\sqrt{C(\alpha)}\epsilon\sqrt{\log(1/\epsilon)}$.

Bounding T_1: This can be bounded in an analogous manner as T_2, thus yielding $T_1 \leq C\sqrt{C(\alpha)\epsilon}\sqrt{\log(2/\epsilon)}$.

Plugging these bounds above, we get

$$\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\|_2 \leq C\sqrt{C(\alpha)\epsilon}\sqrt{\log(1/\epsilon},$$

which proves the statement of the lemma. □

C.2 Proof of Lemma 1

Proof. Consider two Ising models with p vertices. For the first Ising model, consider one edge with parameter 2ϵ. The second Ising model has no edges.

Via a simple calculation, the TV distance between these Ising models can be computed to be $\frac{1}{2}\tanh(2\epsilon) \leq \epsilon$. Consequently, the $\ell_{2,\infty}$ norm of the difference in parameters is ϵ, and this proves the lower bound. □

D Proofs of Propositions in Section 3

D.1 A general result for estimators based on Yatracos classes

Here, we present a result for estimators of the form

$$\mathbb{P}_{\text{est}} = \underset{\mathbb{P}\in\mathcal{P}}{\text{argmin}} \sup_{A\in\mathcal{A}} \left|\mathbb{P}(A) - \widehat{\mathbb{P}}_{n,\epsilon}(A)\right|, \tag{16}$$

where $\widehat{\mathbb{P}}_{n,\epsilon}$ the empirical distribution of n samples from the mixture model $\mathbb{P}_\epsilon$ defined in (1) and $\mathcal{P}$ is the class of all distributions. Recall that $\mathcal{A}$ is defined as

$$\mathcal{A} = \{A(\mathbb{P}_1, \mathbb{P}_2) : \mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}\}, \text{ and } A(\mathbb{P}_1, \mathbb{P}_2) = \{x : \mathbb{P}_1(x) > \mathbb{P}_2(x)\}$$

The result in formally stated in Proposition 2.

Proposition 9. *Given n samples from the mixture model $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}^\star + \epsilon Q$, the estimator $\mathbb{P}_{est}$ defined in (16) satisfies*

$$d_{\text{TV}}(\mathbb{P}_{\text{est}}, \mathbb{P}^\star) \leq 2\epsilon + 2 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_\epsilon(x)\right|$$

Proof. We begin by using $2d_{\text{TV}}(\mathbb{P}_{\text{est}}, \mathbb{P}^\star) = \sum_{x\in\mathcal{X}} |\mathbb{P}_{\text{est}}(x) - \mathbb{P}^\star(x)|$. Consider the sets $B = \{x : \mathbb{P}_{\text{est}}(x) > \mathbb{P}^\star(x)\}$ and $C = \{x : \mathbb{P}_{\text{est}}(x) \leq \mathbb{P}^\star(x)\}$.

This gives us:

$$\begin{aligned}
\sum_{x\in\mathcal{X}} |\mathbb{P}_{\text{est}}(x) - \mathbb{P}^\star(x)| &= 2 \max_{A\in\{B,C\}} \left|\sum_{x\in A} \mathbb{P}_{\text{est}}(x) - \mathbb{P}^\star(x)\right| \\
&\leq 2 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \mathbb{P}_{\text{est}}(x) - \sum_{x\in A} \mathbb{P}^\star(x)\right| \\
&= 2 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \mathbb{P}_{\text{est}}(x) - \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) + \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}^\star(x)\right| \\
&\leq 2 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \mathbb{P}_{\text{est}}(x) - \sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x)\right| + 2 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}^\star(x)\right| \\
&\overset{(i)}{\leq} 4 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}^\star(x)\right| \\
&= 4 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_\epsilon(x) + \sum_{x\in A} \mathbb{P}_\epsilon(x) - \sum_{x\in A} \mathbb{P}^\star(x)\right| \\
&\leq 4 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_\epsilon(x)\right| + 4 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \mathbb{P}_\epsilon(x) - \sum_{x\in A} \mathbb{P}^\star(x)\right| \\
&= 4 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_\epsilon(x)\right| + 4d_{\text{TV}}(\mathbb{P}_\epsilon, \mathbb{P}^\star) \\
&\overset{(ii)}{\leq} 4 \sup_{A\in\mathcal{A}} \left|\sum_{x\in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x\in A} \mathbb{P}_\epsilon(x)\right| + 4\epsilon,
\end{aligned}$$

where in step (i) we have used the optimality of $\mathbb{P}_{\text{est}}$ and in step (ii) we have used the fact that $d_{\text{TV}}(\mathbb{P}_\epsilon, \mathbb{P}^\star) \leq \epsilon$ and this completes the proof. □

D.2 Proof of Lemma 2

With the general result for estimators based on Yatracos classes, we state the proof of Lemma 2.

Proof. For the estimator in (7), the class of distributions is $\mathcal{G}_{p,k}$. Via Proposition 9, we have that:

$$d_{\mathrm{TV}}(\mathbb{P}_{\widehat{\theta}}, \mathbb{P}_{\theta^\star}) \leq 2\epsilon + 2 \underbrace{\sup_{A \in \mathcal{A}} \left| \sum_{x \in A} \widehat{\mathbb{P}}_{n,\epsilon}(x) - \sum_{x \in A} \mathbb{P}_{\epsilon}(x) \right|}_{T_1}$$

Note that distributions in $\mathcal{G}_{p,k}$ are Ising model distributions and are parameterized. Thus, we can alternatively identify the sets $A(\mathbb{P}_1, \mathbb{P}_2)$ via the parameters of Ising model distributions as $A(\theta^{(1)}, \theta^{(2)})$.

Bounding T_1: The set $A(\theta^{(1)}, \theta^{(2)})$ is equivalent to

$$A(\theta^{(1)}, \theta^{(2)}) = \{x : \log \mathbb{P}_{\theta^{(1)}}(x) > \log \mathbb{P}_{\theta^{(2)}}(x)\}$$

Recalling the definitions of $\mathbb{P}_{\theta^{(1)}}$ and $\mathbb{P}_{\theta^{(2)}}$, and flattening the parameters to $\mathbb{R}^{\binom{p}{2}}$, we have:

$$A(\theta^{(1)}, \theta^{(2)}) = \left\{ y : \left\langle \theta^{(1)}_{\text{flat}} - \theta^{(2)}_{\text{flat}}, y \right\rangle + \log(Z(\theta^{(2)})) - \log(Z(\theta^{(1)})) > 0 \right\} = \{y : \langle w, \widetilde{y} \rangle > 0\}$$

where $w = [\theta^{(1)}_{\text{flat}} - \theta^{(2)}_{\text{flat}}, \log(Z(\theta^{(2)})) - \log(Z(\theta^{(1)}))]$ and $\tilde{y} = [y, 1]$. $Z(\theta)$ is the normalization constant of the probability mass function of an Ising model $\mathbb{P}_\theta$ and $y \in \mathbb{R}^{\binom{p}{2}}$ is a vector of sufficient statistics. Since $\theta^{(1)}, \theta^{(2)} \in \mathcal{G}_{p,k}$, both $\theta^{(1)}_{\text{flat}}$ and $\theta^{(2)}_{\text{flat}}$ can have at most k entries. Consequently, the vector w can have at most $2k+1$ non-zero entries. Hence, $\mathcal{A}$ can be viewed as a collection of sets:

$$\mathcal{A} = \{\mathbb{I}\{\langle w, y \rangle > 0\} : w \in \mathbb{R}^{\binom{p}{2}}, ||w||_0 \leq 2k+1\}$$

The following proposition bounds the VC-dimension of sparse linear classifiers:

Proposition 10 ([1, Corollary 1]). *Consider the class of linear predictors, defined by the set $S_s = \{v : ||v||_0 \leq s, v \in \mathbb{R}^m\}$* i.e. *the set of s-sparse vectors. The VC-dimension of this class is upper bounded as: $O(s \log(^{em}/_{s}))$.*

Therefore, from the above proposition, we have that the VC-dimension of $\mathcal{A}$ is bounded from above by $\mathcal{O}\left(2k+1) \log(^{ep^2}/_{4k+2}\right)$ which is $\mathcal{O}(k \log(^{ep}/_{k}))$. Hence, by a concentration of measure argument, we have that with probability at least $1-\delta$:

$$T_1 \lesssim \sqrt{\frac{k \log(^{ep}/_{k})}{n}} + \sqrt{\frac{\log(^{1}/_{\delta})}{n}}.$$

Finally, we obtain

$$d_{\mathrm{TV}}(\mathbb{P}_{\widehat{\theta}}, \mathbb{P}_{\theta^\star}) \leq 2\epsilon + \mathcal{O}\left(\sqrt{\frac{k \log(^{ep}/_{k})}{n}} + \sqrt{\frac{\log(^{1}/_{\delta})}{n}} \right),$$

and this recovers the statement of the lemma. □

E Proof of Propositions in Section 4

E.1 Proof of Theorem 2

Proof Sketch. We give an outline of the proof of the theorem. $\mathbb{P}_{\theta^\star}$ is an Ising model in the high temperature regime with constant α. Recall the proposed estimator:

$$\widehat{\theta}(i) = \operatorname*{argmin}_{w \in \mathcal{N}_d^{\gamma}(\mathcal{S}^{p-2})} \sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(w; x^{(j)})\}_{j=1}^n \right) \right|. \tag{17}$$

Proposition 5 states that the negative conditional log-likelihood $\mathcal{L}_{\theta^\star,i}$ is $C_2(\alpha)$-strongly convex, where $C_2(\alpha)$ is a universal constant only depending on α. Therefore, by the monotonicity of the gradient of strongly-convex function, we bound the parameter error $\|\widehat{\theta}(i) - \theta^\star(i)\|_2$ as

$$\|\widehat{\theta}(i) - \theta^\star(i)\|_2^2 \le \frac{1}{C_2(\alpha)} \left\langle \nabla \mathcal{L}_{\theta^\star,i}(\widehat{\theta}(i)) - \nabla \mathcal{L}_{\theta^\star,i}(\theta^\star(i)), \widehat{\theta}(i) - \theta^\star(i) \right\rangle.$$

Next, note that

$$\begin{aligned}
\|\widehat{\theta}(i) - \theta^\star(i)\|_2 &\le \frac{1}{C_2(\alpha)} \frac{\left\langle \nabla \mathcal{L}_{\theta^\star,i}(\widehat{\theta}(i)) - \nabla \mathcal{L}_{\theta^\star,i}(\theta^\star(i)), \widehat{\theta}(i) - \theta^\star(i) \right\rangle}{\|\widehat{\theta}(i) - \theta^\star(i)\|_2} \\
&\overset{(i)}{\le} \frac{1}{C_2(\alpha)} \sup_{u \in \mathcal{N}_{2d}(\mathcal{S}^{p-2})} \left| \left\langle u, \nabla \mathcal{L}_{\theta^\star,i}(\widehat{\theta}(i)) \right\rangle \right| \\
&\overset{(ii)}{\le} \frac{2}{C_2(\alpha)} \sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \left\langle u, \nabla \mathcal{L}_{\theta^\star,i}(\widehat{\theta}(i)) \right\rangle \right|,
\end{aligned}$$

where in step (i) we have used the facts that 1) $\frac{\widehat{\theta}(i) - \theta^\star(i)}{\|\widehat{\theta}(i) - \theta^\star(i)\|_2}$ is a unit vector with at most $2d$ non-zero elements, and 2) $\nabla \mathcal{L}_{\theta^\star,i}(\theta^\star(i)) = \mathbf{0}$ by the maximum likelihood principle, and in step (ii) we have constructed a $1/2$-cover of the set $\mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})$.

We further analyze the right hand side by splitting it into two different terms as follows.

$$\begin{aligned}
&\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \left\langle u, \mathcal{L}_{\theta^\star,i}(\widehat{\theta}(i)) \right\rangle \right| \le \\
&\qquad \underbrace{\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \left\langle u, \mathcal{L}_{\theta^\star,i}(\widehat{\theta}(i)) \right\rangle - \mathsf{1DMean}\left(\{u^T \nabla \ell_i(\widehat{\theta}(i), x^{(j)})\}_{j=1}^n \right) \right|}_{T_1} \\
&\qquad\qquad + \underbrace{\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(\widehat{\theta}(i), x^{(j)})\}_{j=1}^n \right) \right|}_{T_2}.
\end{aligned}$$

In Lemmas 11 and 12, considering $\gamma = \max\left\{\frac{\epsilon}{p}, \frac{\log(1/\delta)}{np}\right\}$, and for sufficiently large n (10), we bound T_1 and T_2 as $T_1 \le \sqrt{C(\alpha)}\left\{\epsilon\sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d\log(p)}{n}} + \sqrt{\frac{d}{n}\log\left(\frac{3ep}{d\gamma}\right)}\right\}$, and in Lemma 12, we bound T_2 as $T_2 \le \sqrt{C(\alpha)}\left\{\epsilon\sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d\log(p)}{n}} + \sqrt{\frac{d}{n}\log\left(\frac{3ep}{d\gamma}\right)}\right\} + \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right)$ respectively.

Plugging these bound into the previous right hand side, we obtain

$$\|\widehat{\theta}(i) - \theta^\star(i)\|_2 \lesssim \sqrt{C(\alpha)} \left\{ \epsilon \sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d\log(p)}{n}} + \sqrt{\frac{d}{n}\log\left(\frac{3ep}{d\gamma}\right)} \right\} + \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right),$$

and this recovers the statement of the theorem. □

We state Lemmas 11 and 12 and prove them below.

Lemma 11. *Consider samples $\{x^{(j)}\}_{j=1}^n$ from the mixture model $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}_{\theta^\star} + \epsilon Q$, where $\mathbb{P}_{\theta^\star}$ is an Ising model over p nodes in the high temperature regime (3) with constant α and with maximum vertex degree d. Suppose n, confidence δ and contamination level ϵ satisfy (10). Then,* 1DMean *satisfies*

$$\sup_{w \in \mathcal{N}_d^\gamma(\mathcal{S}^{p-2})} \sup_{u \in \mathcal{N}_d^{1/2}(\mathcal{S}^{p-2})} \left| \langle u, \nabla \mathcal{L}_{\theta^\star,i}(w) \rangle - \mathsf{1DMean}\left(\{u^T \nabla \ell_i(w; x^{(j)})\}_{j=1}^n \right) \right| \\ \leq \sqrt{C(\alpha)} \left\{ \epsilon \sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d\log(p)}{n}} + \sqrt{\frac{d}{n}\log\left(\frac{3ep}{d\gamma}\right)} \right\}.$$

Proof. Let $z \sim \mathbb{P}_{\theta^\star}$. In the proof of Lemma 7, we showed that

$$\Pr(|\langle u, \nabla \ell_i(w; z)\rangle|) \leq 2\exp\left(-\frac{t^2}{C(\alpha)}\right)$$

holds due to the form of the gradient and the sub-Gaussianity of the Ising model distribution. This implies that the gradients of ℓ_i due to non-outlier samples are sub-Gaussian. This allows us to leverage techniques from [40] to produce a guarantee for the 1DMean algorithm when the true distribution is sub-Gaussian in Lemma 13. This states that

$$\left| \langle u, \nabla \mathcal{L}_{\theta^\star,i}(w) \rangle - \mathsf{1DMean}\left(\{u^T \nabla \ell_i(w; x^{(j)})\}_{j=1}^n \right) \right| \lesssim \epsilon \sqrt{C(\alpha)\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{C(\alpha)}{n}\log\left(\frac{1}{\delta}\right)},$$

where $w \in \mathcal{N}_d^\gamma(\mathcal{S}^{p-2})$ and $u \in \mathcal{N}_d^{1/2}(\mathcal{S}^{p-2})$.

Finally, to convert the point-wise bound to a uniform bound, we perform a union bound over all the elements in $\mathcal{N}_d^\gamma(\mathcal{S}^{p-2})$ and $\mathcal{N}_d^{1/2}(\mathcal{S}^{p-2})$, and use the fact that the number of elements in the cover can be bounded as $|\mathcal{N}_k^\gamma(\mathcal{S}^{p-2})| \leq \left(\frac{3ep}{k\gamma}\right)^k$ to recover the statement of the result. □

Lemma 12. *Given samples $\{x^{(j)}\}_{j=1}^n$ from the mixture model $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}_{\theta^\star} + \epsilon\mathbb{Q}$, where $\mathbb{P}_{\theta^\star}$ is an Ising model over p nodes in the high temperature regime (3) with constant α, there exists a constant $C(\alpha)$ that only depends on α such that:*

$$\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(\widehat{\theta}(i); x^{(j)})\}_{j=1}^n \right) \right| \\ \leq \sqrt{C(\alpha)} \left\{ \epsilon \sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d\log(p)}{n}} + \sqrt{\frac{d}{n}\log\left(\frac{3ep}{d\gamma}\right)} \right\} + \max\left(\epsilon, \frac{\log(1/\delta)}{n}\right)$$

where $\widehat{\theta}(i)$ is as defined in (9) with $\gamma = \max\left\{\frac{\epsilon}{p}, \frac{\log(1/\delta)}{p}\right\}$.

Proof. First, define $C_\gamma(\theta^\star(i))$ as the element closest to $\theta^\star(i)$ in the set $\mathcal{N}_d^\gamma(\mathcal{S}^{p-2})$.

We begin the proof by recognizing that

$$\begin{aligned}
&\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(\widehat{\theta}(i); x^{(j)})\}_{j=1}^n \right) \right| \\
&\qquad\qquad \overset{(i)}{\leq} \sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(C_\gamma(\theta^\star(i)); x^{(j)})\}_{j=1}^n \right) \right| \\
&\overset{(ii)}{\leq} \underbrace{\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(C_\gamma(\theta^\star(i)); x^{(j)})\}_{j=1}^n \right) - \langle u, \nabla \mathcal{L}_{\theta^\star,i}(C_\gamma(\theta^\star(i))) \rangle \right|}_{T_{2,1}} \\
&\qquad\qquad\qquad\qquad + \underbrace{\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \langle u, \nabla \mathcal{L}_{\theta^\star,i}(C_\gamma(\theta^\star(i))) \rangle \right|}_{T_{2,2}}
\end{aligned}$$

where Step (i) uses the optimality of $\widehat{\theta}(i)$ and Step (ii) performs splitting by addition and subtraction as mentioned earlier.

Bounding $T_{2,1}$: $T_{2,1}$ can be bounded using Lemma 11, since it holds for any $w \in \mathcal{N}_d^\gamma(\mathcal{S}^{p-2})$ and $C_\gamma(\theta^\star(i)) \in \mathcal{N}_d^\gamma(\mathcal{S}^{p-2}$ by definition. Therefore, we get

$$\begin{aligned}
&\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(C_\gamma(\theta^\star(i)); x^{(j)})\}_{j=1}^n \right) - \langle u, \nabla \mathcal{L}_{\theta^\star,i}(C_\gamma(\theta^\star(i))) \rangle \right| \\
&\qquad\qquad \leq \sqrt{C(\alpha)} \left\{ \epsilon \sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d \log(p)}{n}} + \sqrt{\frac{d}{n} \log\left(\frac{3ep}{d\gamma}\right)} \right\}.
\end{aligned}$$

Bounding $T_{2,2}$: $T_{2,2}$ can be bounded as follows:

$$\begin{aligned}
\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \langle u, \nabla \mathcal{L}_{\theta^\star,i}(C_\gamma(\theta^\star(i))) \rangle \right| &\leq \| \nabla \mathcal{L}_{\theta^\star,i}(C_\gamma(\theta^\star(i))) \|_2 \\
&= \| \nabla \mathcal{L}_{\theta^\star,i}(C_\gamma(\theta^\star(i))) - \nabla \mathcal{L}_{\theta^\star,i}(\theta^\star(i)) \|_2 \\
&\leq L \| C_\gamma(\theta^\star(i)) - \theta^\star(i) \|_2 \leq L\gamma,
\end{aligned}$$

where L is the Lipschitz constant of $\mathcal{L}_{\theta^\star,i}$. A simple calculation reveals that:

$$\begin{aligned}
\nabla^2 \mathcal{L}_{\theta^\star,i}(w) &= \mathbb{E}_{x \sim P_{\theta^\star}} [\sigma(\langle w, x(-i) \rangle x_i)(1 - \sigma(\langle w, x(-i) \rangle x_i)) x(-i) x(-i)^T] \\
\Rightarrow v^T \nabla^2 \mathcal{L}_{\theta^\star,i}(w) v &= \mathbb{E}_{x \sim P_{\theta^*}} [\sigma(\langle w, x(-i) \rangle x_i)(1 - \sigma(\langle w, x(-i) \rangle x_i))(\langle v, x(-i) \rangle)^2] \\
&\overset{(i)}{\leq} \frac{1}{4} \mathbb{E}_{x \sim P_{\theta^*}} [(v^T x_i)^2] \overset{(ii)}{\leq} \frac{p}{4} \|v\|_2^2
\end{aligned}$$

where in Step (i) we have used the fact that $\sigma(z)(1 - \sigma(z)) \leq \frac{1}{4}$ and in Step (ii) we have used the Cauchy-Schwarz inequality, leading to $L = p$.

With the choice of $\gamma = \max\left\{ \frac{\epsilon}{p}, \frac{\log(1/\delta)}{n} \right\}$, we have the final result

$$\begin{aligned}
&\sup_{u \in \mathcal{N}_{2d}^{1/2}(\mathcal{S}^{p-2})} \left| \mathsf{1DMean}\left(\{u^T \nabla \ell_i(\widehat{\theta}(i); x^{(j)})\}_{j=1}^n \right) \right| \\
&\qquad \leq \sqrt{C(\alpha)} \left\{ \epsilon \sqrt{\log\left(\frac{1}{\epsilon}\right)} + \sqrt{\frac{d \log(p)}{n}} + \sqrt{\frac{d}{n} \log\left(\frac{3ep}{d\gamma}\right)} \right\} + \max\left(\epsilon, \frac{\log(1/\delta)}{n} \right)
\end{aligned}$$

and this completes the proof. □

E.1.1 Auxiliary Results

Here we state and prove Lemma 13, which we use in the proof of Lemma 11.

Lemma 13 ([40, Lemma 3])**.** *Suppose $\mathbb{P}^\star$ is a sub-Gaussian distribution with variance proxy σ^2 and mean $\mu = \mathbb{E}_{x\sim\mathbb{P}^\star}[x]$. Given n samples from the mixture distribution $\mathbb{P}_\epsilon = (1-\epsilon)\mathbb{P}^\star + \epsilon Q$, Algorithm 1 returns an estimate $\widehat{\theta}_\delta$ that satisfies*

$$|\widehat{\theta}_\delta - \mu| \lesssim \epsilon\sqrt{\sigma^2 \log\left(\frac{1}{\epsilon}\right)} + \sqrt{\sigma^2 \log\left(\frac{1/\delta}{n}\right)}$$

with probability at least $1-\delta$.

Proof. The proof mostly follows the proof in [40].

Let $I^\star$ be the interval $\mu \pm \sqrt{\sigma^2 \log\left(\frac{1}{\delta_1}\right)}$. For notational convenience, let $f_n(u,v) = \sqrt{u(1-u)}\sqrt{\frac{\log(1/v)}{n}} + \frac{2}{3}\frac{\log(1/v)}{n}$. Let $\widehat{I} = [a,b]$ be the interval obtained using the first split of the sample set $\mathcal{Z}_1$ *i.e.* the shortest interval containing $n(1-(\delta_1+\epsilon+f_n(\epsilon+\delta_1,\delta_3)))$ points of $\mathcal{Z}_1$. In Algorithm 1, we have $\delta_1 = \epsilon$ and $\delta_3 = \delta/4$.

From [40, Claim 5], we have that

$$\text{length}(\widehat{I}) \leq \text{length}(I^\star) \leq 2\sqrt{\sigma^2 \log\left(\frac{1}{\delta_1}\right)}.$$

To bound the error of the estimator, we analyze the quantity

$$\left|\frac{1}{|\widehat{I}|}\sum_{z_i\in\mathcal{Z}_2} z_i \mathbb{I}\left\{z_i \in \widehat{I}\right\} - \mu\right|,$$

where $|\widehat{I}| = \sum\limits_{z_i\in\mathcal{Z}_2} \mathbb{I}\left\{z_i \in \widehat{I}\right\}$.

We do so by casing on whether a sample z_i was sampled from $\mathbb{P}^\star$ or from Q, like so.

$$\begin{aligned}\left|\frac{1}{|\widehat{I}|}\sum_{z_i\in\mathcal{Z}_2} z_i \mathbb{I}\left\{z_i \in \widehat{I}\right\} - \mu\right| &= \left|\frac{1}{|\widehat{I}|}\left(\sum_{\substack{z_i\in\mathcal{Z}_2\\ z_i\sim\mathbb{P}^\star}} z_i \mathbb{I}\left\{z_i \in \widehat{I}\right\} + \sum_{\substack{z_i\in\mathcal{Z}_2\\ z_i\sim Q}} z_i \mathbb{I}\left\{z_i \in \widehat{I}\right\}\right) - \mu\right| \\ &\leq \underbrace{\left|\frac{1}{|\widehat{I}|}\sum_{\substack{z_i\in\mathcal{Z}_2\\ z_i\sim\mathbb{P}^\star}} z_i \mathbb{I}\left\{z_i \in \widehat{I}\right\} - \mu\right|}_{T_1} + \underbrace{\left|\frac{1}{|\widehat{I}|}\sum_{\substack{z_i\in\mathcal{Z}_2\\ z_i\sim Q}} z_i \mathbb{I}\left\{z_i \in \widehat{I}\right\} - \mu\right|}_{T_2}.\end{aligned}$$

Bounding T_1: From [40, Claim 6], we bound T_1 with probability at least $1-\delta_3-\delta_5$ as

$$T_1 \leq \frac{\epsilon + f_n(\epsilon,\delta_5)}{1-\delta_4}\cdot 4\sqrt{\sigma^2 \log\left(\frac{1}{\delta_1}\right)},$$

where $\delta_4 = (\delta_1+\epsilon) + f_n(\delta_1+\epsilon,\delta_3)$.

Bounding T_2: To bound T_2, we split the terms further.

$$
\begin{aligned}
T_2 = \left| \frac{1}{|\widehat{I}|} \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \in \widehat{I} \\ z_i \sim Q}} (z_i - \mu) \right| &= \frac{|\widehat{I}_{\mathbb{P}^\star}|}{|\widehat{I}|} \left| \frac{1}{|\widehat{I}_{\mathbb{P}^\star}|} \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \in \widehat{I} \\ z_i \sim Q}} (z_i - \mu) \right| \\
&\leq \underbrace{\frac{|\widehat{I}_{\mathbb{P}^\star}|}{|\widehat{I}|} \left| \left(\frac{1}{|\widehat{I}_{\mathbb{P}^\star}|} \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \in \widehat{I} \\ z_i \sim Q}} z_i \right) - \mathbb{E}[x | x \in \widehat{I}, x \sim \mathbb{P}^\star] \right|}_{T_{2,1}} \\
&\quad + \underbrace{\frac{|\widehat{I}_{\mathbb{P}^\star}|}{|\widehat{I}|} \left| \mathbb{E}[x | x \in \widehat{I}, x \sim \mathbb{P}^\star] - \mu \right|}_{T_{2,2}},
\end{aligned}
$$

where $|\widehat{I}_{\mathbb{P}^\star}| = \sum_{\substack{z_i \in \mathcal{Z}_2 \\ z_i \sim \mathbb{P}^\star}} \mathbb{I}\left\{ z_i \in \widehat{I} \right\}$ is the number of elements in $\mathcal{Z}_2$ that were originally sampled from $\mathbb{P}^\star$.

$T_{2,1}$ is the deviation of the mean of the samples originally sampled from Q and remain in $\widehat{I}$ from the mean of $\mathbb{P}^\star$ conditioned on the event that they belong to $\widehat{I}$ as well. $T_{2,2}$ measures the deviation of the mean of $\mathbb{P}^\star$ from the mean of the same distribution conditioned on $\widehat{I}$.

Bounding $T_{2,1}$: We bound $T_{2,1}$ using [40, Lemma 15]. With this result, we get that with probability at least $1 - \delta_7$,

$$
T_{2,1} \leq \sqrt{\frac{2\sigma^2 \log(3/\delta_7)}{\mathbb{P}^\star(\widehat{I})}} + 2\sqrt{\sigma^2 \log\left(\frac{1}{\delta_1}\right)} \frac{\log(3/\delta_7)}{|\widehat{I}_{\mathbb{P}^\star}|}.
$$

Bounding $T_{2,2}$: To control $T_{2,2}$ we make use of Proposition 8 in conjuction with [40, Lemma 14] to get

$$
T_{2,2} \leq 2\mathbb{P}^\star(\widehat{I}^c) \sqrt{\sigma^2 \log\left(\frac{1}{\mathbb{P}^\star(\widehat{I}^c)}\right)},
$$

where $\mathbb{P}^\star(A)$ is the probability that $z \sim \mathbb{P}^\star$ lies in A. Finally, we bound $\mathbb{P}^\star(\widehat{I}^c$ using [40, Claim 7] to obtain with probability at least $1 - \delta_6$ that

$$
\mathbb{P}^\star(\widehat{I}^c) \leq C_1\epsilon + C_2\delta_1 + C_3\frac{\log(n)}{n} + C_4\frac{\log(1/\delta_6)}{n} + C_5\frac{\log(1/\delta_3)}{n},
$$

where $\{C_i\}_{i=1}^6$ are universal constants.

Therefore, combining the bounds for T_1, $T_{2,1}$ and $T_{2,2}$, and setting $\delta_1 = \epsilon$, $\delta_3 = \delta_5 = \delta_6 = \delta_7 = \delta/4$ and noting that for the choice of n $|\widehat{I}_{\mathbb{P}^\star}| \geq \frac{n}{2}$, we get the final deviation bound:

$$
T_1 + T_{2,1} + T_{2,2} \lesssim \epsilon\sqrt{\sigma^2 \log\left(\frac{1}{\epsilon}\right)} + \sqrt{\sigma^2 \log\left(\frac{1/\delta}{n}\right)},
$$

and this completes the proof of the lemma. □

F Proof of Theorem 3

In this section, we present the proof of Theorem F. The proof mostly follows the analysis in the proofs of Lemma 7 and Theorem 1. The only difference is that we will not be able to use the sub-Gaussianity of Ising model distributions anymore, as it is no longer applicable.

Proof. Following the proof of Lemma 7, we have for any v such that $\|v\|_1 = 1$ that

$$\begin{aligned}\left|\left\langle v, \nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| &\leq \left|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_1)\right\rangle \mathbb{I}\left\{z_1 \neq z_2\right\}\right]\right| \\ &\quad + \left|\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_2)\right\rangle \mathbb{I}\left\{z_1 \neq z_2\right\}\right]\right| \\ &\overset{(i)}{\leq} \underbrace{\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left|\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_1)\right\rangle\right| \mathbb{I}\left\{z_1 \neq z_2\right\}\right]}_{T_1} \\ &\quad + \underbrace{\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left|\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_2)\right\rangle\right| \mathbb{I}\left\{z_1 \neq z_2\right\}\right]}_{T_2},\end{aligned}$$

where in step (i), we have used Jensen's inequality for $f(x) = |x|$.

Bounding T_1: By Hölder's inequality $\left|\left\langle v, \nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| \leq \|v\|_1 \left\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\|_\infty$. Again by Jensen's inequality, and the explicit form of $\nabla \ell_i$, we have $\left\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\|_\infty = \left\|\mathbb{E}\left[\nabla \ell_i(2\theta^{(2)}(i), z_1)\right]\right\|_\infty \leq \mathbb{E}\left[\|\nabla \ell_i(2\theta^{(2)}(i), z_1)\|_\infty\right] \leq 1$. Therefore,

$$\mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\left|\left\langle v, \nabla \ell_i(2\theta^{(2)}(i); z_1)\right\rangle\right| \mathbb{I}\left\{z_1 \neq z_2\right\}\right] \leq \mathbb{E}_{z_1,z_2\sim\mathcal{C}}\left[\mathbb{I}\left\{z_1 \neq z_2\right\}\right] \leq \epsilon.$$

Bounding T_2: T_2 can be bounded in the exact same way as T_1.

Plugging these bounds, we get that

$$\left|\left\langle v, \nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| \leq 2\epsilon.$$

Now, following the first part of the proof of Theorem 1, we have using Hölder's inequality and the bound above that

$$\frac{C}{2}\frac{\exp(-O(\omega))}{\omega}\|\varDelta_i\|_2^2 \leq \left|\left\langle \varDelta_i, \nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\rangle\right| \leq \|\varDelta_i\|_1 \left\|\nabla \mathcal{L}_{\theta^{(1)},i}(2\theta^{(2)}(i))\right\|_\infty \leq 2\epsilon\|\varDelta_i\|_1$$

where $\varDelta_i = 2\theta^{(1)}(i) - 2\theta^{(2)}(i)$. Now, since $\theta^{(1)}$ and $\theta^{(2)}$ are parameters of Ising models with maximum vertex degree d, $\varDelta_i = 2\theta^{(1)}(i) - 2\theta^{(2)}(i)$ has atmost $2d$ non-zero elements. Consequently, we get $\|\varDelta_i\|_1 \leq \sqrt{d}\|\varDelta_i\|_2$.

Finally, plugging the above norm inequality in the previous bound, we have:

$$\|\varDelta_i\|_2 \lesssim \epsilon\sqrt{d}\omega\exp(O(\omega)).$$

Analogously, since $d \leq k$ when $G^{(1)}, G^{(2)} \in \mathcal{G}_{p,k}$, we have

$$\|\varDelta_i\|_2 \lesssim \epsilon\sqrt{k}\omega\exp(O(\omega)),$$

Alternatively, note that by the triangle inequality: $\|\varDelta_i\|_1 \leq \|2\theta^{(1)}(i)\|_1 + \|2\theta^{(2)}(i)\|_1 \leq 4\omega$. This gives us:

$$\|\varDelta_i\|_2 \lesssim \sqrt{\epsilon}\omega\exp(O(\omega))$$

Since both types of inequalities holds simultaneously, we recover the statements of the theorem for $\mathcal{G}_{p,d}$ and $\mathcal{G}_{p,k}$. □