Proceedings of Machine Learning Research vol 134:1–38, 2021 34rd Annual Conference on Learning Theory

Outlier-Robust Learning of Ising Models Under Dobrushin's Condition

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin, Madison
Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego
Alistair Stewart STEWART.AL@GMAIL.COM
Web 3 Foundation
Yuxin Sun YXSUN@CS.WISC.EDU
University of Wisconsin, Madison

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study the problem of learning Ising models satisfying Dobrushin's condition in the outlier-robust setting where a constant fraction of the samples are adversarially corrupted. Our main result is to provide the first computationally efficient robust learning algorithm for this problem with near-optimal error guarantees. Our algorithm can be seen as a special case of an algorithm for robustly learning a distribution from a general exponential family. To prove its correctness for Ising models, we establish new anti-concentration results for degree-2 polynomials of Ising models that may be of independent interest.

Keywords: Robust High-Dimensional Statistics, Ising models

1. Introduction

1.1. Background and Motivation

Probabilistic graphical models (Koller and Friedman, 2009) provide a rich and unifying framework to model structured high-dimensional distributions in terms of the local dependencies between the input variables. The problem of inference in graphical models arises in many applications across scientific disciplines, see, e.g., Wainwright and Jordan (2008). In this work, we study the inverse problem of learning graphical models from data. Various formalizations of this general learning problem have been studied during the past five decades, see, e.g., Chow and Liu (1968); Dasgupta (1997); Abbeel et al. (2006); Wainwright et al. (2006); Anandkumar et al. (2012); Santhanam and Wainwright (2012); Loh and Wainwright (2012); Bresler et al. (2013, 2014); Bresler (2015); Klivans and Meka (2017)), resulting in general theory and algorithms for various settings.

In this work, we focus on learning Ising models (Ising, 1925), the prototypical family of binary undirected graphical models with applications in computer vision, computational biology, and statistical physics (Li, 2009; Jaimovich et al., 2006; Felsenstein, 2004; Chatterjee, 2005).

Definition 1 (Ising Model) *Given a symmetric matrix $(\theta_{ij})_{i,j\in[d]}$ with zero diagonal and a vector $(\theta_i)_{i\in[d]}$, the Ising model distribution P_θ is defined as follows: For any $x \in \{\pm 1\}^d$, $P_\theta(x) = \frac{1}{Z(\theta)} \exp\left((1/2)\sum_{i,j\in[d]} \theta_{ij}x_ix_j + \sum_{i=1}^d \theta_i x_i\right)$, where the normalizing factor $Z(\theta)$ is called the partition function. We call the matrix $(\theta_{ij})_{i,j\in[d]} \in \mathbb{R}^{d\times d}$ the interaction matrix and the vector $(\theta_i)_{i\in[d]} \in \mathbb{R}^d$ the external field.*

The majority of prior algorithmic work on learning Ising models studies the "structure learning" problem, i.e., the problem of learning the structure of the underlying graph of non-zero entries of the interaction matrix, see, e.g., Bresler (2015); Klivans and Meka (2017); Hamilton et al. (2017). In this line of work, it is assumed that the true graph satisfies some structural property (typically, a tree or bounded-degree structure) and certain (upper and lower) bounds are imposed on the underlying parameters. Such assumptions are information-theoretically necessary for this version of the problem. An emerging line of work studies the *distribution learning* problem, i.e., the task of computing an Ising model that is close to the target in total variation distance, see, e.g., Dagan et al. (2020); Daskalakis and Pan (2020); Bhattacharyya et al. (2020) for a few recent papers.

Here we study the algorithmic problem of learning Ising models in the presence of *adversarially corrupted data*. We focus on the following standard data corruption model that generalizes Huber's contamination model (Huber, 1964).

Definition 2 (Total Variation Contamination) *Given $0 < \epsilon < 1/2$ and a class of distributions $\mathcal{F}$ on $\mathbb{R}^d$, the* adversary *operates as follows: The algorithm specifies the number of samples n. The adversary knows the true target distribution $X \in \mathcal{F}$ and selects a distribution F such that $d_{\mathrm{TV}}(F, X) \leq \epsilon$. Then n i.i.d. samples are drawn from F and are given as input to the algorithm.*

Intuitively, the parameter ϵ in Definition 2 quantifies the power of the adversary. The total variation contamination model is strictly stronger than Huber's contamination model. Recall that in Huber's model (Huber, 1964), the adversary generates samples from a mixture distribution F of the form $F = (1-\epsilon)X + \epsilon N$, where X is the unknown target distribution and N is an adversarially chosen noise distribution. That is, in Huber's model the adversary is only allowed to add outliers.

The contamination setting we consider is standard in robust statistics (Hampel et al., 1986; Huber and Ronchetti, 2009), a field which seeks to develop *outlier-robust* estimators — algorithms that can tolerate a *constant fraction* of corrupted datapoints, independent of the dimension. Classical work, starting with Tukey and Huber in the 60s, developed statistically optimal robust estimators for a number of settings. However, these early methods lead to exponential time algorithms even for the most basic high-dimensional estimation tasks (e.g., mean estimation).

A recent line of work, starting with Diakonikolas et al. (2016); Lai et al. (2016), has developed the first computationally efficient and outlier-robust learning algorithms for a range of "simple" high-dimensional probabilistic models. Since these initial algorithmic works, we have witnessed substantial research progress on algorithmic aspects of robust high-dimensional estimation by several communities, see, e.g., Diakonikolas and Kane (2019) for a recent survey on the topic.

Prior algorithmic work on learning graphical models has almost exclusively studied the uncontaminated setting, where the data are i.i.d. samples from the distribution of interest. Some recent work (Hamilton et al., 2017; Goel et al., 2019; Katiyar et al., 2020) has developed algorithms for structure learning in the (significantly weaker) *independent failures model*, where the coordinates of each example are independently flipped/missing with some probability. On the other hand, Lindgren et al. (2019) point out that structure learning becomes information-theoretically impossible in the contamination model, if an adversary is allowed to corrupt even a tiny fraction of the samples.

The only algorithmic work we are aware of in the contamination model is by Cheng et al. (2018) who developed an outlier-robust learner for low-degree Bayes nets (directed graphical models) with known graph structure. We also note that very recent work (Prasad et al., 2020) developed nearly tight *sample complexity* bounds for learning Ising models in Huber's contamination model under various structural assumptions — albeit by using underlying estimators that run in exponential time.

1.2. Our Contributions

In this work, we study the following version of the learning problem: *Given a set of corrupted samples from an unknown Ising model, our goal is to learn the underlying distribution in total variation distance.* This is a natural (and standard) formulation of distribution learning that has been studied extensively, even in the uncontaminated setting. *Our main result is the first computationally efficient outlier-robust estimator for Ising models* in this setting, under some natural assumptions. We note that we do not make structural assumptions about the underlying graph — our algorithms work with Ising models on the complete graph.

To state our contributions in detail, we require some additional terminology.

Definition 3 (Dobrushin's condition) *Given an Ising model P_θ with interaction matrix $(\theta_{ij})_{i,j\in[d]}$ and external field $(\theta_i)_{i\in[d]}$, we say that it satisfies Dobrushin's condition if $\max_{i\in[d]} \sum_{j\neq i} |\theta_{ij}| \leq 1-\eta$ for some constant $0<\eta<1$.*

Dobrushin's condition for Ising models is a classical assumption needed to rule out certain pathological behaviors. It is standard in various areas, including statistical physics, machine learning, and theoretical CS (Külske, 2003; Götze et al., 2019; Dagan et al., 2020; Adamczak et al., 2019; Gheissari et al., 2018; Marton, 2015).

Our main result is an efficient algorithm for outlier-robust learning of Ising models with zero external field satisfying Dobrushin's condition.

Theorem 4 (Robustly Learning Ising Models) *Let $X\sim P_{\theta^*}$ be an Ising model without external field satisfying Dobrushin's condition for some universal constant $\eta>0$. There is a universal constant $\epsilon_0>0$ such that the following holds: Let $0<\epsilon<\epsilon_0$ and S' be an ϵ-corrupted set of N samples from P_{θ^*}. There is a $\mathrm{poly}(N,d)$ time algorithm that, for some $N=\widetilde{O}_\eta(d^2/\epsilon^2)$, on input S' and ϵ, returns a symmetric matrix $\widehat{\theta}\in\mathbb{R}^{d\times d}$ such that with probability at least $99/100$, we have that $\|\widehat{\theta}-\theta^*\|_F\leq O_\eta(\epsilon\log(1/\epsilon))$. Moreover, the Ising model distribution $P_{\widehat{\theta}}$ satisfies Dobrushin's condition and $d_{\mathrm{TV}}(P_{\widehat{\theta}},P_{\theta^*})\leq O_\eta(\|\widehat{\theta}-\theta^*\|_F)\leq O_\eta(\epsilon\log(1/\epsilon))$.*

Some comments are in order. We note that any estimator information-theoretically requires error $\Omega(\epsilon)$ in the contamination model. That is, the error guarantee of our algorithm is optimal, within logarithmic factors. Moreover, our algorithm is proper (i.e., it outputs an Ising model) and performs parameter learning, i.e., it estimates the desired parameters to sufficient accuracy to yield the desired total variation distance guarantee.

Our techniques extend to yield an outlier-robust learning algorithm with the same error guarantee for Ising models with non-zero external field (under additional assumptions). Due to space limitations, these extensions are deferred to Appendix C.3. For the non-zero external field case, the value $\widehat{\theta}$ that we recover unfortunately is not guaranteed to be close to θ^* in Frobenius norm. In fact, this is the wrong norm to compare them in and such an approximation is information-theoretically impossible. However, we do still guarantee that the corresponding Ising model distribution $P_{\widehat{\theta}}$ satisfies Dobrushin's condition and $d_{\mathrm{TV}}(P_{\widehat{\theta}},P_{\theta^*})\leq O(\epsilon\log(1/\epsilon))$. Intuitively , as long as the dependencies among each point and the external fields are sufficiently small, we can robustly learn the Ising model distribution in total variation distance.

To achieve our results, we view the Ising model as an instance of a general exponential family.

Definition 5 (Exponential Family) *An exponential family in canonical form is a family of distributions P_θ supported on a set $\mathcal{X}$, where the parameter θ belongs to some convex set $\Omega \subseteq \mathbb{R}^d$, with density function $P_\theta(x) = \exp\left(\langle T(x), \theta\rangle - A(\theta)\right), \forall x \in \mathcal{X}$, where $A(\theta)$ is the normalizing factor called log-partition function and the vector $T(x)$ is called the sufficient statistics of P_θ.*

As one of our main contributions, we provide a computationally efficient outlier-robust parameter learning algorithm for exponential families under the following condition.

Condition 6 *For an arbitrary $\theta \in \Omega$, the exponential family P_θ satisfies the following:*

1. $\mathbf{Cov}_{X\sim P_\theta}[T(X)] \succeq c_1\, I$, *where $c_1 > 0$ is a universal constant independent of θ and the dimension d of $T(X)$.*
2. *$T(x)$ has sub-exponential tails for a universal constant $c_2 > 0$, i.e., for any unit vector $v \in \mathbb{R}^d$, it holds that $\mathbf{Pr}_{X\sim P_\theta}[|\langle v, T(X) - \mathbf{E}[T(X)]\rangle| > t] \leq 2\exp(-c_2 t)$, for all $t > 0$, where $c_2 > 0$ is a universal constant independent of θ and the dimension d of $T(X)$.*
3. *There is an algorithm that, given as input $\theta \in \Omega$ and $\gamma > 0$, it runs in $\mathrm{poly}(d, 1/\gamma)$ time and it outputs i.i.d. samples from a distribution D_γ such that $d_{\mathrm{TV}}(D_\gamma, P_\theta) \leq \gamma$.*

In addition, the diameter of Ω is bounded above and we can efficiently compute approximate projections on Ω. Specifically, it holds that $\mathrm{diam}(\Omega) \leq \exp(\mathrm{poly}(d))$, and for any $\delta > 0$ and $z \in \mathbb{R}^d$, there is a $\mathrm{poly}(d, 1/\delta)$ time algorithm that computes a point $y \in \Omega$ such that $\|y - P_\Omega(z)\|_2 \leq \delta$, where P_Ω is the projection operation.

Theorem 7 (Robust Learning of Exponential Families) *Let P_{θ^*} be an exponential family over $\mathcal{X}$ with sufficient statistics $T(x)$, where the parameter $\theta^* \in \Omega$ and $\Omega \subseteq \mathbb{R}^d$ is convex. Assume that Condition 6 holds. Let $0 < \epsilon < \epsilon_0$, for some universal constant ϵ_0, and S' be an ϵ-corrupted set of N samples from P_{θ^*}. There is a $\mathrm{poly}(N, d)$ time algorithm that, for some $N = \widetilde{O}(d/\epsilon^2)$, on input S' and $\epsilon > 0$, returns a vector $\widehat{\theta} \in \Omega$ such that with probability at least $99/100$ we have that $\|\widehat{\theta} - \theta^*\|_2 \leq O(\epsilon \log(1/\epsilon))$. In addition, $d_{\mathrm{TV}}(P_{\widehat{\theta}}, P_{\theta^*}) \leq O(\|\widehat{\theta} - \theta^*\|_2) \leq O(\epsilon \log(1/\epsilon))$.*

As we will explain in the next subsection, our robust learning algorithm for Ising models is the algorithm given in Theorem 7. The main technical challenge is in establishing correctness, i.e., showing that an Ising model under Dobrushin's condition satisfies Condition 6. To achieve this, we develop new anti-concentration results for degree-2 polynomial of Ising models that we believe may be of independent interest (see Theorem 19).

1.3. Overview of Techniques

Our outlier-robust learning algorithm for Ising models is a special case of a robust learning algorithm for the class of exponential families (satisfying Condition 6). We start with an intuitive description of this algorithm followed by a brief sketch of the tools required to prove its correctness.

To robustly learn a family of distributions in total variation distance, one typically requires a set of relevant parameters and a "parameter distance", so that sufficiently accurate approximation in parameter distance implies approximation in total variation distance. For exponential families, a natural set of parameters present themselves: the expectation of the sufficient statistics of the distribution. Our strategy will be to robustly estimate this expectation.

Unfortunately, there is a wrinkle in this strategy which relates to the scale in which we are working. On the one hand, in order to robustly estimate the mean of a distribution, one needs to know some sort of tail bounds on the set of clean samples; and for these tail bounds to hold, we need to know the scale at which we expect this decay to happen. On the other hand, once we learn an approximation to the true mean of the sufficient statistics, we need to relate the sizes of these errors to the errors we will obtain in the underlying parameters for the family, and to the total variation distance of the final distribution that we learn. These relationships define certain natural scales for our problem, and it is not clear how to obtain a robust algorithm if these scales disagree (in such a case, the accuracy to which we can learn the expectation of the sufficient statistics might differ from the accuracy to which we need to learn it to obtain good error in total variation distance) or if the relevant scale depends on the underlying (unknown) parameters.

To resolve this issue, we need to make an assumption (Condition 6). Specifically, we need to assume that there is a convex set Ω of parameters in our exponential family, such that any elements of the family inside this set have sufficient statistics whose covariances are within constant multiples of each other. This implies that the relevant scales for our problem are all comparable.

From this point, there is a relatively straightforward algorithm that achieves suboptimal error. After a change of variables, we can assume that within Ω all of the sufficient statistics have covariance proportional to the identity. This allows us to use standard robust mean estimation algorithms (Fact 12) to estimate the mean of the sufficient statistics to error $O(\sqrt{\epsilon})$ in ℓ_2-norm. This in turn allows us to estimate our distribution to error $O(\sqrt{\epsilon})$ in total variation distance.

To improve on this error guarantee, we will need to obtain better error in our robust mean estimation algorithm. This can be achieved under the following assumptions: (1) The distribution in question satisfies strong tail bounds. (2) We know an accurate approximation to the covariance matrix of the distribution. As for (1), it follows for general exponential families that their sufficient statistics will have exponential tail bounds, which is sufficient for us. For (2), we will need to already have a good approximation of the underlying parameters of our distribution. This gives rise to an iterative algorithm. If we know the underlying parameters of our exponential family to error δ, we can learn the mean of the sufficient statistics — and thus new approximations to the parameters — to error $O(\epsilon \log(1/\epsilon) + \sqrt{\delta\epsilon})$ (Lemma 15). Iterating this several times, we can eventually achieve the near-optimal error of $O(\epsilon \log(1/\epsilon))$.

Our result for Ising models is obtained via an application of the above algorithm. Ising models are a special case of an exponential family, where the sufficient statistics are given by degree-2 polynomials. For the above algorithm to provably work, we need to show that (under some reasonable conditions on parameters) the covariance of the sufficient statistics is well-behaved. In particular, we show that if the underlying parameters satisfy the Dobrushin condition, the covariance matrix of the sufficient statistics will be proportional to the identity.

Interestingly, Dagan et al. (2020) recently showed that this holds for the covariance of the space of *degree*-1 polynomials of such Ising models. We need to generalize this to show that $\mathbf{Var}[X^T A X]$ is proportional to $\|A\|_F^2$ for any symmetric matrix A with zero diagonal. To achieve this, we use a decoupling trick to reduce the problem to the degree-1 case. We relate the variance of $X^T A X$ to $\mathbf{E}[|(X+Y)^T A (X-Y)|^2]$, for X and Y independent copies of our distribution. If we condition on the set S of coordinates where $X_i = Y_i$, then $(X+Y)$ and $(X-Y)$ become independent Ising models. By estimating the covariances of these *linear* functions of these statistics, we can get a handle on the final bound.

Organization After some technical preliminaries (Section 2), in Section 3 we prove Theorem 7. In Section 4, we establish Theorem 4. Due to space limitations, our results for the non-zero external field and several technical proofs have been deferred to the Appendix.

2. Preliminaries

Notation For $d \in \mathbb{Z}_+$, we use $[d]$ to denote the set $\{1, \ldots, d\}$. Given a subset $S \subseteq [d]$, we will denote $-S = [d] \setminus S$. In particular, given $i \in [d]$, let $-i = [d] \setminus \{i\}$. Given a vector $a = (a_1, \ldots, a_d)$ and $S \subseteq [d]$, let a_S denote the $|S|$-coordinate vector $\{a_i : i \in S\}$. Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ be the d-dimensional unit sphere. Given a real symmetric matrix $A \in \mathbb{R}^{d \times d}$, let $\|A\|_F \stackrel{\text{def}}{=} \sqrt{\sum_{i,j \in [d]} A_{ij}^2}$, let $\|A\|_2 \stackrel{\text{def}}{=} \max_{v \in \mathbb{S}^{d-1}} \|Av\|_2$, and let $\|A\|_\infty \stackrel{\text{def}}{=} \max_{i \in [d]} \sum_{j=1}^d |A_{ij}|$.

For $u, v \in \mathbb{R}^d$, we use $\langle u, v \rangle$ for the inner product of u and v. For any two symmetric matrices $A, B \in \mathbb{R}^{d \times d}$, we say that $A \succeq B$ if $A - B$ is positive semi-definite (PSD), and $A \preceq B$ if $B - A$ is positive semi-definite. For any two distributions p, q over a probability space Ω, let $d_{\mathrm{TV}}(p, q) \stackrel{\text{def}}{=} \sup_{S \subseteq \Omega} |p(S) - q(S)|$ denote the total variation distance between p and q and let $d_{\mathrm{KL}}(p, q) \stackrel{\text{def}}{=} \int_\Omega \log\left(\frac{dp}{dq}\right) dp$ denote the KL-divergence of p and q. We use $\mathbf{E}[X], \mathbf{Var}[X], \mathbf{Cov}[X, Y]$ to denote the expectation of X, variance of X, and covariance of X and Y respectively.

Additional Technical Preliminaries Basic properties of Ising models and exponential families are given in Appendices A.3 and A.7. We will also require sub-exponential distributions and their basic properties (see Appendix A.1). We will use the following terminology.

Definition 8 (Bounded Ising Model) *Given $M, \alpha > 0$, we say that an Ising model distribution P_θ is (M, α)-bounded if $\max_{i \in [d]} \sum_{j \neq i} |\theta_{ij}| \leq M$ and $\max_{i \in [d]} |\theta_i| \leq \alpha$.*

Intuitively, the first inequality states that the dependencies among the points are weak and the second inequality guarantees that the variance of each point is sufficiently large.

Glauber dynamics is the canonical Markov chain for sampling from undirected graphical models. The dynamics on the Ising model defines a reversible, ergodic Markov chain with stationary distribution corresponding to the Ising model. (We describe the dynamics in Appendix A.5.) The Glauber dynamics for an Ising model satisfying Dobrushin's condition is rapidly mixing, i.e., it converges fast to the underlying distribution P_θ.

Fact 9 (see, e.g., Levin and Peres (2017)) *Let P_θ be an Ising model satisfying Dobrushin's condition and $\gamma > 0$. Then, after $t = \Omega(d(\log d + \log(1/\gamma)))$ steps of Glauber dynamics, we have that $d_{\mathrm{TV}}\left(X^{(t)}, P_\theta\right) \leq \gamma$.*

Fact 9 tells us that given the parameter θ, we can efficiently generate approximate random samples from the Ising model distribution P_θ, as long as it satisfies Dobrushin's condition.

Maximum Likelihood Estimation Given a set of i.i.d. samples $S = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ drawn from an exponential family P_θ with sufficient statistics $T(x)$ and unknown parameter $\theta \in \Omega$, the principle of maximum likelihood allows us to compute an estimate $\widehat{\theta} \in \Omega$ by maximizing the likelihood of S, i.e., $l(\theta, S) = \frac{1}{n} \sum_{i=1}^n \ln P_\theta(x_i) = \frac{1}{n} \sum_{i=1}^n (\langle T(x_i), \theta \rangle - A(\theta)) = \langle \theta, \widehat{\mu}_T \rangle - A(\theta)$, where $\widehat{\mu}_T = \frac{1}{n} \sum_{i=1}^n T(x_i)$ is the empirical mean of the sufficient statistics $T(x)$ defined by the

point set S. Define $L(\theta, \mu_T) = \langle \theta, \mu_T \rangle - A(\theta)$ and fix μ_T to be the empirical mean $\widehat{\mu}_T$. The maximum likelihood estimator $\widehat{\theta}$ is chosen to maximize the objective function $L(\theta, \widehat{\mu}_T)$ over $\theta \in \Omega$.

The following lemma states that under suitable conditions, if we obtain a good estimate of the mean μ_T of the sufficient statistics $T(x)$, the maximum likelihood estimator (MLE) will be a good approximation of the parameter θ (see Appendix A.8 for the proof).

Lemma 10 *Let P_{θ^*} be an exponential family such that θ^* lies in a convex set $\Omega \subseteq \mathbb{R}^d$. Let $\mu_T^* = \mathbf{E}_{X \sim P_{\theta^*}}[T(X)]$ and $\Sigma_T^* = \mathbf{Cov}_{X \sim P_{\theta^*}}[T(X)]$. Let μ_T' be an approximation of μ_T^* such that $\|\mu_T' - \mu_T^*\|_2 \leq \delta$. Let $\theta' \in \arg\max_{\theta \in \Omega} L(\theta, \mu_T')$, where $L(\theta, \mu_T') = \langle \theta, \mu_T' \rangle - A(\theta)$. If there is a universal constant $c > 0$ such that $\mathbf{Cov}_{X \sim P_\theta}[T(X)] \succeq c\, I$, for all $\theta \in \Omega$, then $\|\theta' - \theta^*\|_2 \leq 2\delta/c$.*

3. Robust Parameter Learning of Exponential Families

In Section 3.1, we give an efficient algorithm (Lemma 11) that reduces parameter estimation of exponential families to the task of estimating the mean of the sufficient statistics. In Sections 3.2 and 3.3, we describe and analyze our computationally efficient robust parameter learning algorithm for exponential families satisfying Condition 6.

3.1. Learning via Estimating the Mean of Sufficient Statistics

Lemma 11 *Let P_{θ^*} be an exponential family with sufficient statistics $T(x)$, where $\theta^* \in \Omega$ and $\Omega \subseteq \mathbb{R}^d$ is convex. Assume that Condition 6 holds. Let $\mu_T^* = \mathbf{E}_{X \sim P_{\theta^*}}[T(X)]$ and μ_T' be an approximation of μ_T^* such that $\|\mu_T' - \mu_T^*\|_2 \leq \delta$, for some $0 < \delta < 1$ sufficiently small. Let $0 < \zeta < 1$. Then there is a $\mathrm{poly}(d, 1/\delta, 1/\zeta)$ time algorithm that, given input μ_T', δ and ζ, returns a vector $\widehat{\theta} \in \Omega$ such that with probability at least $1 - \zeta$ we have that $\|\widehat{\theta} - \theta^*\|_2 \leq O(\delta)$.*

We give a proof sketch here; the details are in Appendix B.1. Let $\theta' = \arg\max_{\theta \in \Omega} L(\theta, \mu_T')$. By Lemma 10, we know that $\|\theta' - \theta^*\|_2 \leq O(\|\mu_T' - \mu_T^*\|_2) \leq O(\delta)$. In addition, since given any $\theta \in \Omega$ we can efficiently sample from a distribution within small total variation distance of P_θ, we can efficiently approximate the gradient $\nabla_\theta(-L(\theta, \mu_T')) = \mathbf{E}_{X \sim P_\theta}[T(X)] - \mu_T'$. Therefore, we can apply projected gradient descent to efficiently obtain an estimate $\widehat{\theta}$ of θ' with $\|\widehat{\theta} - \theta'\|_2 \leq O(\delta)$, using the fact that $-L(\theta, \mu_T')$ is L-smooth and m-strongly convex for some constants $L, m > 0$.

3.2. Robust Parameter Learning Algorithm

The pseudocode of our algorithm is given in Algorithm 1. We make essential use of the following previously known algorithms for robust mean estimation under bounded and approximately known covariance assumptions.

Fact 12 (Diakonikolas et al. (2017); Steinhardt et al. (2018)) *Let D be a distribution supported on $\mathbb{R}^d$ with unknown mean μ and unknown covariance Σ such that $\Sigma \preceq \sigma^2 I$, for some $\sigma > 0$. Let $0 < \epsilon < \epsilon_0$, for some universal constant ϵ_0, and $\delta = O(\sqrt{\epsilon})$. Given an ϵ-corrupted set of N samples drawn from D, for some $N = \widetilde{O}(d/\epsilon)$, there is a $\mathrm{poly}(N, d)$ time algorithm that outputs a vector $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 \leq O(\sigma\delta) = O(\sigma\sqrt{\epsilon})$ with high probability.*

Fact 13 (see, e.g., Cheng et al. (2019)) *Let D be a distribution on $\mathbb{R}^d$ with unknown mean μ and unknown covariance Σ. Let $0 < \epsilon < \epsilon_0$, for some universal constant ϵ_0, $\tau \leq O(\sqrt{\epsilon})$, and $\delta =$*

$O(\sqrt{\tau\epsilon} + \epsilon\log(1/\epsilon))$. *Suppose that D has sub-exponential tails and Σ satisfies $\|\Sigma - I\|_2 \le \tau$. Given an ϵ-corrupted set of N samples drawn from D, for some $N = \widetilde{O}(d/\epsilon^2)$, there is a* $\mathrm{poly}(N, d)$ *time algorithm that outputs a vector $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 \le O(\delta)$ with high probability.*

Algorithm 1 starts by applying the robust mean estimation routine of Fact 12 and Lemma 11 to obtain an initial estimate $\theta^{(0)}$ with ℓ_2-error $O(\sqrt{\epsilon})$. Starting from this rough estimate, the algorithm applies an iterative refinement procedure (see Fact 13 and Lemma 15) for $T = O(\log\log(1/\epsilon))$ iterations to achieve near-optimal ℓ_2-error of $O(\epsilon\log(1/\epsilon))$.

Algorithm 1: Robust parameter estimation for exponential families

Input : $0 < \epsilon < \epsilon_0$, ϵ-corrupted set of $N = \widetilde{O}(d/\epsilon^2)$ samples from exponential family P_{θ^*} satisfying Condition 6, with $\mu_T^* = \mathbf{E}_{X\sim P_{\theta^*}}[T(X)]$ and $\Sigma_T^* = \mathbf{Cov}_{X\sim P_{\theta^*}}[T(X)]$.

Output: Parameter $\widehat{\theta} \in \mathbb{R}^d$ such that $\|\widehat{\theta} - \theta^*\|_2 \le O(\epsilon\log(1/\epsilon))$ with high probability.

Let $\delta = O(\sqrt{\epsilon})$.
Compute $\widehat{\mu}_T^{(0)}$ with $\|\widehat{\mu}_T^{(0)} - \mu_T^*\|_2 \le \delta$ by applying the robust mean estimation algorithm of Fact 12.
Compute $\theta^{(0)} \in \Omega$ by applying projected gradient descent to the function $-L\big(\theta, \widehat{\mu}_T^{(0)}\big)$.
Let $\tau_0 = O(\delta)$ be an upper bound of $\|\theta^{(0)} - \theta^*\|_2$.
Let $K = O(\log\log(1/\epsilon))$.
for k = 0 *to* $K - 1$ **do**
  Let $n = \widetilde{O}(d^2/\tau_k^2)$ and $X^{(1)}, \ldots, X^{(n)}$ be i.i.d. random samples such that $d_{\mathrm{TV}}(X^{(i)}, P_{\theta^{(k)}}) \le \widetilde{O}(\tau_k^2/d^2)$.
  Let $\mu_T^{(k)} = \frac{1}{n}\sum_{i=1}^n X^{(i)}$ and $\Sigma_T^{(k)} = \frac{1}{n}\sum_{i=1}^n \big(X^{(i)} - \mu_T^{(k)}\big)\big(X^{(i)} - \mu_T^{(k)}\big)^T$.
  Let $\delta = O(\sqrt{\epsilon\tau_k} + \epsilon\log(1/\epsilon))$.
  Compute $\widehat{\mu}$ with $\big\|\widehat{\mu} - \big(\Sigma_T^{(k)}\big)^{-1/2}\mu_T^*\big\|_2 \le \delta$ by applying the robust mean estimation algorithm of Fact 13.
  Compute $\theta^{(k+1)} \in \Omega$ by applying projected gradient descent to the function $-L\big(\theta, \big(\Sigma_T^{(k)}\big)^{1/2}\widehat{\mu}\big)$.
  Let $\tau_{k+1} = O(\delta)$ be an upper bound of $\|\theta^{(k+1)} - \theta^*\|_2$.
end
return $\theta^{(K)}$.

To prove correctness, we require Lemmas 14 and 15 below. Roughly speaking, in each refinement step, we first apply Lemma 14 to obtain a covariance estimate $\Sigma_T^{(k)}$ given the current parameter estimate $\theta^{(k)}$. Then, by Lemma 15, we are able to obtain a more accurate parameter estimate $\theta^{(k+1)}$.

Lemma 14 shows that given an estimate θ' of the true parameter θ^* of the exponential family satisfying Condition 6 with $\|\theta' - \theta^*\|_2 \le \delta$, we can efficiently compute an estimate $\widehat{\Sigma}_T$ of the true covariance Σ_T^* such that $\big\|\widehat{\Sigma}_T - \Sigma_T^*\big\|_2 \le O(\delta)$ with high probability.

Lemma 14 *Let P_θ^* be an exponential family with sufficient statistics $T(x)$, where $\theta^* \in \Omega$ and $\Omega \subseteq \mathbb{R}^d$ is convex. Assume that Condition 6 holds. Let $\Sigma_T^* = \mathbf{Cov}_{X\sim P_{\theta^*}}[T(X)]$. Let $\theta' \in \Omega$ be such that $\|\theta' - \theta^*\|_2 \le \delta$, for some $\delta > 0$. Let $0 < \zeta < 1$. There is a* $\mathrm{poly}(d, 1/\delta, 1/\zeta)$ *algorithm that, given as input θ', δ and ζ, returns a $d \times d$ PSD matrix $\widehat{\Sigma}_T$ such that with probability at least $1 - \zeta$, we have that $\big\|\widehat{\Sigma}_T - \Sigma_T^*\big\|_2 \le O(\delta)$.*

The algorithm establishing Lemma 14 is very simple – it corresponds to lines 7 and 8 of Algorithm 1. Roughly speaking, we first generate i.i.d. random samples from a distribution Q which is close to $P_{\theta'}$, and then let $\widehat{\Sigma}_T$ be the empirical covariance of these samples.

Lemma 15 shows that, given a fairly accurate estimate of the covariance Σ_T^* of an exponential family satisfying Condition 6, we can efficiently obtain a more accurate estimate of θ^*.

Lemma 15 (Iterative Refinement) *Let $0 < \delta < \delta_0$ for some universal constant δ_0 sufficiently small. Let $0 < \zeta < 1$. Assume that Condition 6 holds. Let S' be an ϵ-corrupted set of N samples from P_{θ^*}. There is an algorithm that, for some $N = \widetilde{O}(d/\epsilon^2)$, given S', δ, ζ, and $\Sigma_T^{(k)}$ with $\left\|\Sigma_T^{(k)} - \Sigma_T^*\right\|_2 \le \delta$, it runs in $\mathrm{poly}(N, 1/\delta, 1/\zeta)$-time and outputs $\theta^{(k+1)} \in \Omega$ such that with probability at least $1-\zeta$ it holds that $\|\theta^{(k+1)} - \theta^*\|_2 \le O(\sqrt{\epsilon\delta} + \epsilon\log(1/\epsilon))$.*

The algorithm establishing Lemma 15 corresponds to lines 9 to 12 of Algorithm 1. The main idea is as follows: Let $Y = (\Sigma_T^{(k)})^{-1/2} T(X)$. We can show that the covariance of Y is close to identity and Y is sub-exponential, for some universal constant $c > 0$. Thus, we can apply the robust mean estimation algorithm of Fact 13 to obtain an estimate $\widehat{\mu}$ of the mean of Y. In addition, we can show that $(\Sigma_T^{(k)})^{1/2}\widehat{\mu}$ is a good estimate of $\mu^* = \mathbf{E}_{X\sim P_{\theta^*}}[T(X)]$, and therefore we can apply Lemma 11 to get a new estimate $\theta^{(k+1)}$.

Before we give the proofs of Lemmas 14 and 15, we show how they imply Theorem 7.

Proof [Proof of Theorem 7] Algorithm 1 starts by applying the robust mean estimation algorithm for bounded covariance distributions (Fact 12) to obtain an estimate $\mu_T^{(0)}$ of the true mean $\mu_T^* = \mathbf{E}_{X\sim P_{\theta^*}}[T(X)]$ such that $\left\|\mu_T^{(0)} - \mu_T^*\right\|_2 \le O(\sqrt{\epsilon})$. Then it applies Lemma 11 to obtain an initial estimate $\theta^{(0)}$ of the underlying parameter θ^* with $\|\theta^{(0)} - \theta^*\|_2 \le O(\sqrt{\epsilon})$.

In each refinement step k, assume that we have a current estimate $\theta^{(k)}$ of the true parameter θ^* such that $\|\theta^{(k)} - \theta^*\|_2 \le \tau_k$, for some $\tau_k > 0$. Algorithm 1 first applies the algorithm of Lemma 14 to compute an estimate $\Sigma_T^{(k)}$ of the true covariance Σ_T^* with $\left\|\Sigma_T^{(k)} - \Sigma_T^*\right\|_2 \le O(\|\theta^{(k)} - \theta^*\|_2) \le O(\tau_k)$. Then it applies the algorithm of Lemma 15 to obtain a more accurate estimate $\theta^{(k+1)}$ of the true parameter θ^* such that $\|\theta^{(k+1)} - \theta^*\|_2 \le \tau_{k+1}$, where $\tau_{k+1} = O(\sqrt{\epsilon\tau_k} + \epsilon\log(1/\epsilon))$. After $K = O(\log\log(1/\epsilon))$ iterations, we obtain an estimate $\widehat{\theta} = \theta^{(K)}$ such that $\|\widehat{\theta} - \theta^*\|_2 \le O(\epsilon\log(1/\epsilon))$. To bound the sample complexity and the failure probability, we take $\zeta = 1/\log(1/\epsilon)$ in Lemmas 11, 14 and 15. Therefore, the sample complexity is $N = \widetilde{O}(dK/\epsilon^2) = \widetilde{O}(d/\epsilon^2)$ and the total failure probability is at most $O(K/\log(1/\epsilon)) \le 1/100$ by a union bound. Finally, by Lemma 35, it follows that $d_{\mathrm{TV}}(P_{\widehat{\theta}}, P_{\theta^*}) \le O(\|\widehat{\theta} - \theta^*\|_2) \le O(\epsilon\log(1/\epsilon))$. ■

3.3. Implementing the Iterative Refinement Steps

In this subsection, we prove Lemmas 14 and 15. The following proposition connects the third derivative of the log-partition function $A(\theta)$ of the exponential family P_θ with the third moment of the sufficient statistics $T(x)$ (see Appendix B.2 for the proof).

Proposition 16 *Let P_θ be an exponential family with sufficient statistics $T(x)$ and density $P_\theta(x) = \exp(\langle T(x), \theta\rangle - A(\theta))$, $\theta \in \mathbb{R}^d$. Let $\mu_T = \mathbf{E}_{X\sim P_\theta}[T(X)]$ and $\Sigma_T = \mathbf{Cov}_{X\sim P_\theta}[T(X)]$. Then, for any $i, j, k \in [d]$, we have that $\frac{\partial(\Sigma_T)_{ij}}{\partial\theta_k} = \mathbf{E}_{X\sim P_\theta}[(T(X) - \mu_T)_i (T(X) - \mu_T)_j (T(X) - \mu_T)_k]$.*

As a consequence of Proposition 16, we can bound the difference between the covariance matrices of the sufficient statistics of two exponential families with sub-exponential tails in terms of the difference between their parameters (see Appendix B.3 for the proof).

Lemma 17 *Let $\Omega \subseteq \mathbb{R}^d$ be a convex set. Assume that for any $\theta \in \Omega$, the exponential family P_θ with sufficient statistics $T(x)$ has sub-exponential tails for a universal constant $c > 0$, i.e., for any $\theta \in \Omega$ and any unit vector $v \in \mathbb{R}^d$, $\mathbf{Pr}_{X\sim P_\theta}[|\langle v, T(X) - \mathbf{E}_{X\sim P_\theta}[T(X)]\rangle| > t] \leq 2\exp(-ct)$. Then there is a constant $c' > 0$ such that for any $\theta^1, \theta^2 \in \Omega$, we have that $\|\Sigma_T(\theta^1) - \Sigma_T(\theta^2)\|_2 \leq c'\|\theta^1 - \theta^2\|_2$, where for any $\theta \in \Omega$, $\Sigma_T(\theta) = \mathbf{Cov}_{X\sim P_\theta}[T(X)]$.*

Proof [Proof of Lemma 14] Let $\Sigma'_T = \mathbf{Cov}_{X\sim P_{\theta'}}[T(X)]$. From Lemma 17, it follows that $\|\Sigma'_T - \Sigma^*_T\|_2 \leq O(\|\theta' - \theta^*\|_2) = O(\delta)$. In addition, given $\theta' \in \Omega$, we can efficiently sample from a distribution within total variation distance $\gamma = \frac{\delta^2\zeta^2}{2d^2(\log d + \log(12/\zeta))}$ from $P_{\theta'}$.

Since $P_{\theta'}$ is sub-exponential and γ is sufficiently small, by standard properties of sub-exponential distributions and the data processing inequality, it follows that the empirical estimate $\widehat{\Sigma}_T$ satisfies $\|\widehat{\Sigma}_T - \Sigma'_T\|_2 \leq O(\delta)$ with probability at least $1 - \zeta$. (Formally, this follows by picking $t = \log(12/\zeta)$ and $n = \frac{d^2(\log d + \log(12/\zeta))}{\delta^2\zeta}$ in Claim 38.) This implies that $\|\widehat{\Sigma}_T - \Sigma^*_T\|_2 \leq \|\widehat{\Sigma}_T - \Sigma'_T\|_2 + \|\Sigma'_T - \Sigma^*_T\|_2 \leq O(\delta)$, completing the proof. ■

Proof [Proof of Lemma 15] Let $Y = \left(\Sigma_T^{(k)}\right)^{-1/2} T(X)$, $\mu_Y = \mathbf{E}_{X\sim P_{\theta^*}}[Y] = \left(\Sigma_T^{(k)}\right)^{-1/2}\mu_T^*$, and $\Sigma_Y = \mathbf{Cov}_{X\sim P_{\theta^*}}[Y] = \left(\Sigma_T^{(k)}\right)^{-1/2}\Sigma_T^*\left(\Sigma_T^{(k)}\right)^{-1/2}$. From Condition 6 and Fact 21, we know that $c\,I \preceq \Sigma_T^* \preceq c'\,I$ for some universal constants $c' \geq c > 0$. Since $\left\|\Sigma_T^{(k)} - \Sigma_T^*\right\|_2 \leq \delta \leq \delta_0$, we have that $(c - \delta_0)\,I \preceq \Sigma_T^{(k)} \preceq (c' + \delta_0)\,I$ and for any unit vector $v \in \mathbb{S}^{d-1}$, we have that

$$\begin{aligned}\left|v^T(\Sigma_T^{(k)})^{-1/2}(\Sigma_T^{(k)} - \Sigma_T^*)(\Sigma_T^{(k)})^{-1/2}v\right| &\leq \left\|(\Sigma_T^{(k)}) - \Sigma_T^*\right\|_2 \left\|(\Sigma_T^{(k)})^{-1/2}v\right\|_2^2 \\ \leq \left\|(\Sigma_T^{(k)}) - \Sigma_T^*\right\|_2 \left\|(\Sigma_T^{(k)})^{-1/2}\right\|_2^2 &= \left\|(\Sigma_T^{(k)}) - \Sigma_T^*\right\|_2 \left\|(\Sigma_T^{(k)})^{-1}\right\|_2 \leq O(\delta)\,,\end{aligned}$$

which implies that

$$\begin{aligned}1 - O(\delta) = v^T(\Sigma_T^{(k)})^{-1/2}\Sigma_T^{(k)}(\Sigma_T^{(k)})^{-1/2}v - O(\delta) &\leq v^T(\Sigma_T^{(k)})^{-1/2}\Sigma_T^*(\Sigma_T^{(k)})^{-1/2}v \\ \leq v^T(\Sigma_T^{(k)})^{-1/2}\Sigma_T^{(k)}(\Sigma_T^{(k)})^{-1/2}v + O(\delta) &= 1 + O(\delta)\,.\end{aligned}$$

Therefore, we have that $\|\Sigma_Y - I\|_2 = \left\|\left(\Sigma_T^{(k)}\right)^{-1/2}\Sigma_T^*\left(\Sigma_T^{(k)}\right)^{-1/2} - I\right\|_2 \leq O(\delta)$. In addition, since $T(x)$ has sub-exponential tails by Condition 6 and $\Sigma_T^{(k)} \succeq (c - \delta_0)\,I$, we know that Y also has sub-exponential tails. Therefore, we can apply the robust mean estimation algorithm for approximately known covariance distributions (Fact 13) to obtain an estimate $\widehat{\mu}$ of μ_Y such that $\|\widehat{\mu} - \mu_Y\|_2 \leq O(\sqrt{\epsilon\delta} + \epsilon\log(1/\epsilon))$ with probability at least $1 - \zeta/2$. We thus have that

$$\begin{aligned}\left\|\left(\Sigma_T^{(k)}\right)^{1/2}\widehat{\mu} - \mu_T^*\right\|_2 = \left\|\left(\Sigma_T^{(k)}\right)^{1/2}\left(\widehat{\mu} - \left(\Sigma_T^{(k)}\right)^{-1/2}\mu_T^*\right)\right\|_2 &\leq \left\|\Sigma_T^{(k)}\right\|_2^{1/2} \cdot \|\widehat{\mu} - \mu_Y\|_2 \\ \leq \sqrt{c' + \delta_0} \cdot \|\widehat{\mu} - \mu_Y\|_2 &= O(\sqrt{\epsilon\delta} + \epsilon\log(1/\epsilon))\,.\end{aligned}$$

Then we apply Lemma 11 by taking $\left(\left(\Sigma_T^{(k)}\right)^{1/2}\widehat{\mu},\ O\left(\sqrt{\epsilon\delta} + \epsilon\log(1/\epsilon)\right), \zeta/2\right)$ as input, to obtain a vector $\theta^{(k+1)} \in \Omega$ such that $\|\theta^{(k+1)} - \theta^*\|_2 \leq O\left(\left\|\left(\Sigma_T^{(k)}\right)^{1/2}\widehat{\mu} - \mu_T^*\right\|_2\right) \leq O(\sqrt{\epsilon\delta} + \epsilon\log(1/\epsilon))$. This completes the proof. ■

4. Robustly Learning Ising Models

In this section, we prove Theorem 4, giving our efficient robust learning algorithm for Ising models without external field under Dobrushin's condition. Due to space limitations, our analogous algorithmic result for the non-zero external field case is given in Appendix C.3.

Throughout this section, we assume that X is an Ising model satisfying the Dobrushin condition for some *fixed* constant $\eta > 0$. Therefore, we will suppress any possible dependence on η in our asymptotic notation in this section.

For the zero external field case, the density of an Ising model is $P_\theta(x) = \frac{1}{Z(\theta)} \exp((1/2) \sum_{i,j\in[d]} \theta_{ij} x_i x_j)$, where $(\theta_{ij})_{i,j\in[d]}$ is a $d \times d$ real symmetric matrix with zero diagonal and $Z(\theta)$ is the partition function. By definition, P_θ is an exponential family with sufficient statistics $T(x) = (x_i x_j)_{1\le i<j\le d}$ and the projection of $T(x)$ on a fixed direction is $X^T A X$, where $A \in \mathbb{R}^{d\times d}$ is a symmetric matrix with zero diagonal and $\|A\|_F^2 = 1/2$.

As already mentioned, we view the Ising model distribution as an instance of a general exponential family and apply Algorithm 1. The challenge lies in proving correctness. Let Ω be the set of all θ such that P_θ satisfies Dobrushin's condition. We will show that Condition 6 is satisfied for Ω, and therefore Algorithm 1 succeeds in our context.

First note that, by our choice of Ω, its diameter is bounded ($\mathrm{diam}(\Omega) = \mathrm{poly}(d)$), and we can efficiently compute the projection of any point $z \in \mathbb{R}^{d\times(d-1)/2}$. Moreover, by Fact 9, we can efficiently approximately sample from Ising models satisfying Dobrushin's condition.

It remains to verify the first two statement of Condition 6. For the second statement, we need the following sub-exponential concentration inequality for quadratic functions of $(1-\eta, \alpha)$-bounded Ising models. (This inequality will also be needed for the non-zero external field case.)

Lemma 18 *Let $X \sim P_\theta$ be an Ising model satisfying Dobrushin's condition and $\max_{i\in[d]} |\theta_i| \le \alpha$, where $\alpha > 0$ is an absolute constant. Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix with zero diagonal and $b \in \mathbb{R}^d$. For any $x \in \{\pm 1\}^d$, define $f(x) = (x-v)^T A (x-v) + b^T x$, where v satisfies $\|v - \mathbf{E}[X]\|_2 \le \delta$, for some constant $\delta > 0$. Then there is a universal constant $c > 0$ such that $\mathbf{Pr}[|f(X) - \mathbf{E}[f(X)]| > t] \le 2\exp\left(-(ct)/(\|A\|_F^2 + \|b\|_2^2)^{1/2}\right)$.*

Lemma 18 can be derived via machinery developed in Götze et al. (2019) (see Appendix C.1). From Lemma 18, it follows that the sufficient statistics $T(x)$ has sub-exponential tails, for some universal constant $c > 0$.

It remains to verify the first statement of Condition 6, i.e., to show that for any $\theta \in \Omega$ the Ising model distribution P_θ satisfies $\mathbf{Cov}_{X\sim P_\theta}[T(X)] \succeq c' I$, for some universal constant $c' > 0$. Equivalently, it suffices to show that for any unit vector $w \in \mathbb{S}^{d\times(d-1)/2-1}$, it holds

$$w^T \mathbf{Cov}_{X\sim P_\theta}[T(X)] w = \mathbf{Var}_{X\sim P_\theta}[w^T T(X)] \ge c' .$$

We start with some very basic intuition about this statement. Note that in the very special case where $\theta_{ij} = 0, \forall i, j \in [d]$, $X \sim P_\theta$ is the uniform distribution on the hypercube, i.e., its coordinates are independent Rademacher random variables. In this case, it is easy to see that for any symmetric matrix $A \in \mathbb{R}^{d\times d}$ we have that $\mathbf{Var}[X^T A X] = 2\sum_{i\neq j} A_{ij}^2$. Intuitively, for any (M, α)-bounded Ising model (possibly containing a non-zero external field) for some constants $M, \alpha > 0$, the entries of X are *nearly* independent, which allows us to prove the following variance lower bound.

Our result in this context is the following theorem, which may be of independent interest.

Theorem 19 *Let $X \sim P_\theta$ be an (M, α)-bounded Ising model (possibly with non-zero external field), for some constants $M, \alpha > 0$. There is a constant $c(M, \alpha) > 0$ such that for any symmetric matrix $A \in \mathbb{R}^{d\times d}$ with zero diagonal and any $v \in \mathbb{R}^d$, we have that*

$$\mathbf{Var}[(X - v)^T A(X - v)] \geq c(M, \alpha)\|A\|_F^2 .$$

Due to space limitations, the proof of Theorem 19 is deferred to Appendix C.2. Here we provide a brief outline of the proof. By definition, we can write

$$\begin{aligned}\mathbf{Var}[(X - v)^T A(X - v)] &= \frac{1}{2}\mathbf{E}\big[\big((X - v)^T A(X - v) - (Y - v)^T A(Y - v)\big)^2\big] \\ &= \frac{1}{2}\mathbf{E}\big[\big((X - Y)^T A(X + Y - 2v)\big)^2\big] .\end{aligned}$$

Since there are dependencies between each X_i and Y_i, it is not easy to bound from below the expectation of the quadratic form directly. By Lemma 32, we know that $\mathbf{Cov}[X] \succeq c'(M, \alpha)\, I$, for some universal constant $c'(M, \alpha) > 0$. A natural idea is to reduce the original problem to lower bounding the variance of a linear form.

Define the random variables $S = \{i \in [d] \mid X_i = Y_i\}$ and $A^S_{ij} = A_{ij}, \forall i \notin S, j \in [d]$, $W^S_i = \mathbb{I}[i \in S]X_i - v_i, \forall i \in [d]$. The key observation is that conditioning on a fixed set S, the marginal distributions of X_S and X_{-S} are *independent* $(2M, 2\alpha)$-bounded Ising model distributions. In addition, conditioning on a fixed set S, $X - Y$ only depends on X_{-S}, and $X + Y - 2v$ only depends on X_S. Therefore, we can write

$$\begin{aligned}&\mathbf{E}_{(X,Y)}\big[\big((X - Y)^T A(X + Y - 2v)\big)^2 \mid S\big] \\ &= \mathbf{E}_{(X_S,X_{-S})}\big[\big(X_{-S}^T A^S W^S\big)^2 \mid S\big] = \mathbf{E}_{X_S}\big[\mathbf{E}_{X_{-S}}\big[\big(X_{-S}^T A^S W^S\big)^2 \mid X_S, S\big] \mid S\big] \\ &\geq \mathbf{E}_{X_S}\big[\lambda_{\min}\big(\mathbf{E}_{X_{-S}}\big[X_{-S}X_{-S}^T \mid S\big]\big)\|A^S W^S\|_2^2 \mid S\big] \\ &\geq c'(M, \alpha)\mathbf{E}\big[\|A^S W^S\|_2^2 \mid S\big] ,\end{aligned}$$

where $\lambda_{\min}\big(\mathbf{E}_{X_{-S}}\big[X_{-S}X_{-S}^T \mid S\big]\big)$ denotes the minimum eigenvalue of $\mathbf{E}_{X_{-S}}\big[X_{-S}X_{-S}^T \mid S\big]$. Given this, we can express $\mathbf{E}\big[\|A^S W^S\|_2^2 \mid S\big]$ in terms of the variance of a linear form, and apply Lemma 32 again to obtain the desired lower bound.

We are now ready to prove the main result of this section.

Proof [Proof of Theorem 4] Let $\Omega = \{(\theta_{ij})_{1\leq i<j\leq d} \in \mathbb{R}^{d\times(d-1)/2} \mid \max_{i\in[d]} \sum_{j=1}^{i-1} |\theta_{ji}| + \sum_{j=i+1}^{d} |\theta_{ij}| \leq 1 - \eta\}$, where $\eta > 0$ is the constant in Definition 3. Let $\theta \in \Omega$ and P_θ be the corresponding Ising model distribution. By definition, P_θ[1] is an exponential family with sufficient statistics $T(x) = (x_i x_j)_{1\leq i<j\leq d}$. In order to apply Algorithm 1, we check each statement in Condition 6 one by one. By our choice of Ω, we know that $\mathrm{diam}(\Omega) = O(d)$ and we can efficiently compute the projection of any point $z \in \mathbb{R}^{d\times(d-1)/2}$. From Fact 9, we can sample from P_θ within total variation distance γ in time $O(d(\log d + \log(1/\gamma)))$, for any $\gamma > 0$. Therefore, the third statement holds. From Lemma 18, there is a universal constant $c > 0$ such that for any symmetric matrix $A \in \mathbb{R}^{d\times d}$ with zero diagonal and any $t > 0$, we have that $\mathbf{Pr}_{X\sim P_\theta}[|X^T AX - \mathbf{E}[X^T AX]| > t] \leq 2\exp(-(ct)/\|A\|_F)$, which implies the second statement in Condition 6. Moreover, by Theorem 19, we know that there is a universal constant $c' > 0$ such that for any symmetric matrix

1. For simplicity, we also use θ to denote the $d \times d$ symmetric matrix with zero diagonal.

$A \in \mathbb{R}^{d\times d}$ with zero diagonal, we have that $\mathbf{Var}[X^T AX] \geq c'\|A\|_F^2$, which implies the first statement in Condition 6. Therefore, by Theorem 7, we can efficiently obtain an estimate $\widehat{\theta} \in \Omega$ such that $d_{\mathrm{TV}}(P_{\widehat{\theta}}, P_{\theta^*}) \leq O(\|\widehat{\theta} - \theta^*\|_F) \leq O(\epsilon \log(1/\epsilon))$ with probability at least $99/100$. In addition, by our algorithm $\widehat{\theta} \in \Omega$, and thus the output hypothesis satisfies Dobrushin's condition. ■

References

P. Abbeel, D. Koller, and A. Y. Ng. Learning factor graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.*, 7:1743–1788, 2006.

R. Adamczak, M. Kotowski, B. Polaczyk, and M. Strzelecki. A note on concentration for polynomials in the ising model. *Electronic Journal of Probability*, 24, 2019.

A. Anandkumar, D. J. Hsu, F. Huang, and S. Kakade. Learning mixtures of tree graphical models. In *NIPS*, pages 1061–1069, 2012.

A. Bhattacharyya, S. Gayen, E. Price, and N. V. Vinodchandran. Near-optimal learning of tree-structured distributions by chow-liu. *CoRR*, abs/2011.04144, 2020. URL https://arxiv.org/abs/2011.04144.

G. Bresler. Efficiently learning Ising models on arbitrary graphs. In *STOC*, pages 771–782, 2015.

G. Bresler, E. Mossel, and A. Sly. Reconstruction of Markov random fields from samples: Some observations and algorithms. *SIAM J. Comput.*, 42(2):563–578, 2013.

G. Bresler, D. Gamarnik, and D. Shah. Structure learning of antiferromagnetic Ising models. In *NIPS*, pages 2852–2860, 2014.

S. Chatterjee. *Concentration inequalities with exchangeable pairs*. PhD thesis, Stanford University, 2005.

Y. Cheng, I. Diakonikolas, D. Kane, and A. Stewart. Robust learning of fixed-structure bayesian networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 10304–10316, 2018. Full version available at https://arxiv.org/abs/1606.07384.

Y. Cheng, I. Diakonikolas, R. Ge, and D. P. Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory, COLT 2019*, pages 727–757, 2019.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theor.*, 14(3):462–467, 1968.

Y. Dagan, C. Daskalakis, N. Dikkala, and A. V. Kandiros. Estimating ising models from one sample. *arXiv preprint arXiv:2004.09370*, 2020.

S. Dasgupta. The sample complexity of learning fixed-structure Bayesian networks. *Machine Learning*, 29(2-3):165–180, 1997.

C. Daskalakis and Q. Pan. Tree-structured ising models can be learned efficiently. *CoRR*, abs/2010.14864, 2020. URL https://arxiv.org/abs/2010.14864.

C. Daskalakis, N. Dikkala, and G. Kamath. Concentration of multilinear functions of the ising model with applications to network data. *Advances in Neural Information Processing Systems*, 30:12–23, 2017.

I. Diakonikolas and D. M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019. URL http://arxiv.org/abs/1911.05911.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of FOCS'16*, pages 655–664, 2016.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 999–1008, 2017.

J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, 2004.

R. Gheissari, E. Lubetzky, and Y. Peres. Concentration inequalities for polynomials of contracting ising models. *Electronic Communications in Probability*, 23, 2018.

S. Goel, D.l M. Kane, and A. R. Klivans. Learning ising models with independent failures. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019*, volume 99 of *Proceedings of Machine Learning Research*, pages 1449–1469. PMLR, 2019. URL http://proceedings.mlr.press/v99/goel19a.html.

F. Götze, H. Sambale, and A. Sinulis. Higher order concentration for functions of weakly dependent random variables. *Electronic Journal of Probability*, 24, 2019.

L. Hamilton, F.Koehler, and A. Moitra. Information theoretic properties of markov random fields, and their algorithmic applications. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 2463–2472, 2017.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics. The approach based on influence functions*. Wiley New York, 1986.

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

P.J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley New York, 2009.

E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift fur Physik*, 31:253–258, 1925.

A. Jaimovich, G. Elidan, H. Margalit, and N. T. Friedman. Towards an integrated protein-protein interaction network: a relational markov network approach. *J. Comput Biol.*, 13:145–64, 2006.

A. Katiyar, V. Shah, and C. Caramanis. Robust estimation of tree structured ising models. *CoRR*, abs/2006.05601, 2020. URL https://arxiv.org/abs/2006.05601.

A. R. Klivans and R. Meka. Learning graphical models using multiplicative weights. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 343–354. IEEE Computer Society, 2017.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

A. K. Kuchibhotla and A. Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

C. 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.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Proceedings of FOCS'16*, 2016.

D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer Publishing Company, Incorporated, 3rd edition, 2009. ISBN 9781848002784.

E. M. Lindgren, V. Shah, Y. Shen, A. G. Dimakis, and A. Klivans. On robust learning of ising models. In *NeurIPS Workshop on Relational Representation Learning*, 2019.

P. L. Loh and M. J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. In *NIPS*, pages 2096–2104, 2012.

K. Marton. Logarithmic sobolev inequalities in discrete product spaces: a proof by a transportation cost distance. *arXiv preprint arXiv:1507.02803*, 2015.

Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

A. Prasad, V. Srinivasan, S. Balakrishnan, and P. Ravikumar. On learning ising models under huber's contamination model. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

N. P. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Information Theory*, 58(7):4117–4134, 2012.

J. Steinhardt, M. Charikar, and G. Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018*, pages 45:1–45:21, 2018.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

M. J. Wainwright and M. I. Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. pages 1465–1472, 2006.

Appendix

Appendix A. Omitted Technical Preliminaries

A.1. Basic Facts about Sub-exponential Distributions

Here we present basic facts about sub-exponential distributions. The reader is referred to Vershynin (2018).

Definition 20 (Sub-Exponential Distribution) *A distribution D over $\mathbb{R}$ is sub-exponential if there is a constant $c > 0$ such that for any $t > 0$, we have $\mathbf{Pr}_{X\sim D}[|X - \mathbf{E}[X]| > t] \leq 2\exp(-c\,t)$. We say that a distribution D' over $\mathbb{R}^d$ is sub-exponential if there is a constant $c' > 0$ such that for any unit vector $v \in \mathbb{S}^{d-1}$ and any $t > 0$, we have that $\mathbf{Pr}_{X\sim D'}[|\langle v, X - \mathbf{E}[X]\rangle| > t] \leq 2\exp(-c'\,t)$.*

Fact 21 *Let X be a mean-zero random variable, and suppose that there is a constant $K > 0$ such that for any $t > 0$, $\mathbf{Pr}[|X| > t] \leq 2\exp(-t/K)$. Then there is a constant $C > 0$ such that for any real number $p \geq 1$, $\mathbf{E}[|X|^p] \leq (CKp)^p$. In addition, there is a constant $C' > 0$ such that for any $0 < |\lambda| < 1/(C'K)$, we have that $\mathbf{E}[\exp(\lambda X)] \leq \exp(C'^2K^2\lambda^2)$.*

The following result establishes that, for any sub-exponential distribution, the empirical mean and empirical covariance converge fast to the true mean and covariance.

Lemma 22 (see, e.g., Vershynin (2018); Kuchibhotla and Chakrabortty (2018)) *Let D be a sub-exponential distribution over $\mathbb{R}^d$ with mean μ and covariance Σ. Let $X_1, \ldots, X_n$ be i.i.d. samples drawn from D, $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the empirical mean, and $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)(X_i - \widehat{\mu}_n)^T$ be the empirical covariance. Then there exist constants $c_1, c_2 > 0$ such that the following holds:*

1. *With probability at least $1 - 2\exp(-t^2)$, we have that*

$$\|\widehat{\mu}_n - \mu\|_2 \leq c_1 \max(\delta, \delta^2)\ ,$$

where $\delta = \sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}$, and

2. *With probability at least $1 - 6\exp(-t)$, we have that*

$$\left\|\widehat{\Sigma}_n - \Sigma\right\|_2 \leq c_2 d\left(\sqrt{\frac{t + \log d}{n}} + \frac{((t + \log d)\log n)^2}{n}\right)\ ,$$

A.2. Basic Facts on Optimization of Smooth and Strongly Convex Functions

In this section, we provide some background on smooth and strongly convex optimization.

Definition 23 *Let $\Omega \subseteq \mathbb{R}^d$ be a convex set and $f : \Omega \to \mathbb{R}$ be twice continuously differentiable. For $m > 0$, we say that f is m-strongly convex over Ω if $\nabla^2 f(x) \succeq mI$, for all $x \in \Omega$. We say that f is L-smooth over Ω if $-LI \preceq \nabla^2 f(x) \preceq LI$ for all $x \in \Omega$.*

Algorithm 2: Projected gradient descent for strongly convex smooth optimization

Input : an m-strongly convex and L-smooth function f over a convex set Ω and a constant $\delta > 0$.

Output: an $\widehat{x} \in \Omega$ such that $\|\widehat{x} - x^*\|_2 \leq \delta$, where $x^* = \arg\min_{x\in\Omega} f(x)$.

Let $x^0 \in \Omega$ be an arbitrary initial point and $T = O\left(\frac{L}{m} \log\left(\frac{\mathrm{diam}(\Omega)}{\delta}\right)\right)$.

for $t = 0$ *to* $T-1$ **do**

  $r^t = x^t - \frac{1}{L}\nabla f(x^t)$.

  $x^{t+1} = \arg\min_{x\in\Omega} \|x - r^t\|_2$.

end

return x^T.

Notation Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set. We denote $\mathrm{diam}(\mathcal{X})$ to be the diameter of $\mathcal{X}$ in Euclidean norm, i.e., $\mathrm{diam}(\mathcal{X}) = \sup_{x,y\in\mathcal{X}} \|x - y\|_2$. For an arbitrary point $x \in \mathbb{R}^d$, we denote $P_{\mathcal{X}}(x)$ to be the Euclidean projection of x to $\mathcal{X}$, i.e., $P_{\mathcal{X}}(x) = \arg\min_{z\in\mathcal{X}} \|z - x\|_2$.

The following projected gradient descent method for minimizing a smooth and strongly convex function is standard.

Fact 24 (Nesterov (2018)) *Let $f : \Omega \to \mathbb{R}$ be L-smooth and m-strongly convex. Let $x^* = \arg\min_{x\in\Omega} f(x)$. The iterates in Algorithm 2 satisfy*

$$\|x^{t+1} - x^*\|_2^2 \leq \left(1 - \frac{m}{L}\right) \|x^t - x^*\|_2^2 .$$

Therefore, after $T = O\left(\frac{L}{m} \log\left(\frac{\mathrm{diam}(\Omega)}{\delta}\right)\right)$ iterations, we have that $\|x^T - x^*\|_2 \leq \delta$.

A.3. Basic Properties of Ising Models

Here we present some basic properties of Ising models, which will be used throughout this paper. Our first property states that if we arbitrarily fix the states of an arbitrary set of points, the conditional distribution of other points is still an Ising model.

Fact 25 *Let $X \sim P_\theta$ be an Ising model supported on $\{\pm 1\}^d$ and $I \subseteq [d]$. For any fixed vector $x_{-I} \in \{\pm 1\}^{-I}$, the conditional distribution of X_I over $\{\pm 1\}^I$ conditioning on $X_{-I} = x_{-I}$ is an Ising model with interaction matrix $\theta'_{ij} = \theta_{ij}$, for all $i, j \in I$, and external field $\theta'_i = \theta_i + \sum_{j\notin I} \theta_{ij} x_j$, for all $i \in I$.*

The proof of this fact is standard, but we provide it here for completeness.

Proof Let $x_I, x'_I \in \{\pm 1\}^I$. We calculate the ratio of conditional probabilities for two configurations x_I and x'_I, as follows:

$$\frac{\mathbf{Pr}[X_I = x_I \mid X_{-I} = x_{-I}]}{\mathbf{Pr}[X_I = x'_I \mid X_{-I} = x_{-I}]} = \frac{\exp\left(\sum_{i,j\in I} \theta_{ij} x_i x_j + \sum_{i\in I} x_i \left(\theta_i + \sum_{j\notin I} \theta_{ij} x_j\right)\right)}{\exp\left(\sum_{i,j\in I} \theta_{ij} x'_i x'_j + \sum_{i\in I} x'_i \left(\theta_i + \sum_{j\notin I} \theta_{ij} x_j\right)\right)}.$$

Therefore, the conditional distribution of X_I conditioning on $X_{-I} = x_{-I}$ is an Ising model with interaction matrix $(\theta_{ij})_{i,j\in I}$ and external field $\theta'_i = \theta_i + \sum_{j\notin I} \theta_{ij} x_j$. ■

Our second property states that for an arbitrary (M, α)-bounded Ising model, every point has sufficiently large variance.

Fact 26 *Let $X \sim P_\theta$ be an (M, α)-bounded Ising model supported on $\{\pm 1\}^d$. Then, for every $i \in [d]$ and $x_i \in \{\pm 1\}$, we have that*

$$\frac{\exp(-2(\alpha + M))}{1 + \exp(-2(\alpha + M))} \leq \mathbf{Pr}[X_i = x_i] \leq \frac{\exp(2(\alpha + M))}{1 + \exp(2(\alpha + M))} \, .$$

Therefore, we also have that

$$\mathbf{Var}[X_i] = 4 \, \mathbf{Pr}[X_i = 1] \, \mathbf{Pr}[X_i = -1] \geq 4 \left(\frac{\exp(-2(\alpha + M))}{1 + \exp(-2(\alpha + M))} \right)^2 \, .$$

Proof By definition of the Ising model, we can write

$$\begin{aligned}
\mathbf{Pr}[X_i = x_i] &= \sum_{x_{-i} \in \{\pm 1\}^{d-1}} \mathbf{Pr}[X_{-i} = x_{-i}] \cdot \mathbf{Pr}[X_i = x_i \mid X_{-i} = x_{-i}] \\
&= \sum_{x_{-i} \in \{\pm 1\}^{d-1}} \mathbf{Pr}[X_{-i} = x_{-i}] \cdot \frac{\exp\left(\theta_i x_i + x_i \sum_{j \neq i} \theta_{ij} x_j\right)}{\exp\left(\theta_i x_i + x_i \sum_{j \neq i} \theta_{ij} x_j\right) + \exp\left(-\theta_i x_i - x_i \sum_{j \neq i} \theta_{ij} x_j\right)} \\
&= \sum_{x_{-i} \in \{\pm 1\}^{d-1}} \mathbf{Pr}[X_{-i} = x_{-i}] \cdot \frac{\exp\left(2\theta_i x_i + 2x_i \sum_{j \neq i} \theta_{ij} x_j\right)}{1 + \exp\left(2\theta_i x_i + 2x_i \sum_{j \neq i} \theta_{ij} x_j\right)} \, .
\end{aligned}$$

Since X is an (M, α)-bounded Ising model and the function $f(t) = \frac{e^t}{1+e^t}$ is monotonically increasing, we have that

$$\frac{\exp(-2(\alpha + M))}{1 + \exp(-2(\alpha + M))} \leq \frac{\exp\left(2\theta_i x_i + 2x_i \sum_{j \neq i} \theta_{ij} x_j\right)}{1 + \exp\left(2\theta_i x_i + 2x_i \sum_{j \neq i} \theta_{ij} x_j\right)} \leq \frac{\exp(2(\alpha + M))}{1 + \exp(2(\alpha + M))} \, , \tag{1}$$

which implies that $\frac{\exp(-2(\alpha+M))}{1+\exp(-2(\alpha+M))} \leq \mathbf{Pr}[X_i = x_i] \leq \frac{\exp(2(\alpha+M))}{1+\exp(2(\alpha+M))}$.

Let $p_i = \mathbf{Pr}[X_i = 1]$. We directly calculate $\mathbf{E}[X_i]$ and $\mathbf{Var}[X_i]$ as follows.

$$\begin{aligned}
\mathbf{E}[X_i] &= \mathbf{Pr}[X_i = 1] - \mathbf{Pr}[X_i = -1] = 2p_i - 1, \\
\mathbf{Var}[X_i] &= 1 - \mathbf{E}[X_i]^2 = 1 - (2p_i - 1)^2 = 4p_i(1 - p_i) \, .
\end{aligned}$$

Hence, from inequality (1), we have that

$$\mathbf{Var}[X_i] = 4 \, \mathbf{Pr}[X_i = 1] \, \mathbf{Pr}[X_i = -1] \geq 4 \left(\frac{\exp(-2(\alpha + M))}{1 + \exp(-2(\alpha + M))} \right)^2 \, ,$$

which completes the proof. ■

A.4. Dobrushin's uniqueness condition

Here we introduce the original definition of Dobrushin's condition through the influence between points in general graphical model.

Definition 27 (Influence in graphical models) *Let D be a distribution over some set of points V. Let S_j denote the set of state pairs (X, Y) which differ only at point j. Then the influence of point $j \in V$ on point $i \in V$ is defined as*

$$I(j,i) = \max_{(X,Y)\in S_j} d_{\mathrm{TV}}(D_i(\cdot \mid X_{-i}), D_i(\cdot \mid Y_{-i})) \ ,$$

where $D_i(\cdot \mid X_{-i}), D_i(\cdot \mid Y_{-i})$ denote the marginal distribution of point i conditioning on X_{-i} and Y_{-i} respectively.

Definition 28 (Dobrushin's uniqueness condition) *Let D be a distribution over some set of points V. Then D is said to satisfy Dobrushin's uniqueness condition if $\max_{i\in V} \sum_{j\in V} I(j,i) < 1$.*

For Ising models, Chatterjee (2005) proves that $\max_{i\in V} \sum_{j\neq i} |\theta_{ij}| < 1$ implies the Dobrushin's uniqueness condition.

A.5. Glauber Dynamics

The Glauber dynamics for Ising models proceeds as follows:

1. Start at any initial state $X^{(0)} \in \{\pm 1\}^d$.
2. Pick a point $i \in [d]$ uniformly at random and update $X_i^{(t)}$ as follows:

$$X_i^{(t+1)} = x \quad \text{w.p.} \quad \frac{\exp\left(\theta_i x + \sum_{j\neq i} \theta_{ij} X_j^{(t)} x\right)}{\exp\left(\theta_i + \sum_{j\neq i} \theta_{ij} X_j^{(t)}\right) + \exp\left(-\theta_i - \sum_{j\neq i} \theta_{ij} X_j^{(t)}\right)} \ .$$

A.6. Concentration and Anti-concentration of Ising models

Several recent works have studied the concentration and anti-concentration of functions of Ising models Gheissari et al. (2018); Götze et al. (2019); Daskalakis et al. (2017); Adamczak et al. (2019). Here we record some results which will be used throughout this article.

The following fact states that for any $(1-\eta, \alpha)$-bounded Ising model, for some constants $\eta, \alpha > 0$, the corresponding Ising model distribution is sub-Gaussian.

Fact 29 (Götze et al. (2019)) *Let P_θ be an Ising model satisfying Dobrushin's condition, and $\max_{i\in[d]} |\theta_i| \leq \alpha$ for some constant $\alpha > 0$. Then there is a constant $c(\alpha, \eta) > 0$ such that for any $b \in \mathbb{R}^d$ and any $t > 0$, we have that*

$$\mathbf{Pr}_{X\sim P_\theta}\left[\left|b^T X - \mathbf{E}\left[b^T X\right]\right| > t\right] \leq 2\exp\left(-\frac{t^2}{c(\alpha,\eta)\|b\|_2^2}\right) \ ,$$

where $\eta > 0$ is the constant in Definition 3.

The following concentration property for quadratic forms of Ising models will be used to establish appropriate concentration inequalities.

Fact 30 (Gheissari et al. (2018)) *Let $X \sim P_\theta$ be an Ising model satisfying Dobrushin's condition. Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix with zero diagonal and $b \in \mathbb{R}^d$. For any $x \in \{\pm 1\}^d$, define $f(x) = (x-v)^T A\,(x-v) + b^T x$, where $v = \mathbf{E}[X]$. Then there is a constant $c(\eta) > 0$ such that*

$$\mathbf{Var}[f(X)] \le c(\eta)(\|A\|_F^2 + \|b\|_2^2)\ ,$$

where η is the constant in Definition 3.

We will require the following fact, which states that if the Ising model satisfies Dobrushin's condition, then changing the state of a single point will have small influence on other ones.

Fact 31 (Dagan et al. (2020)) *Let P_θ be an Ising model satisfying Dobrushin's condition. Fix $i \in [d]$ and let μ^1_{-i} denote the conditional expectation over x_{-i} conditioning on $x_i = 1$, and μ^{-1}_{-i} denote the conditional expectation over x_{-i} conditioning on $x_i = -1$. Then, we have that $\left\|\mu^1_{-i} - \mu^{-1}_{-i}\right\|_1 \le 2(1-\eta)/\eta$, and $\sum_{j\neq i} |\mathbf{Cov}[X_i, X_j]| \le (1-\eta)/\eta$, where $\eta > 0$ is the constant in Definition 3.*

We will also require the following anti-concentration result for linear forms on bounded Ising models:

Fact 32 (Dagan et al. (2020)) *Let $X \sim P_\theta$ be an (M,α)-bounded Ising model, where $M, \alpha > 0$ are constants. Then there is a constant $c(M,\alpha) > 0$ such that for any vector $b \in \mathbb{R}^d$, we have that*

$$\mathbf{Var}[b^T X] \ge c(M,\alpha)\|b\|_2^2\ .$$

As a consequence of Fact 32, for any (M,α)-bounded Ising model X, we have that $\mathbf{Cov}[X] \succeq c(M,\alpha)\, I$.

A.7. Basic Properties of Exponential Families

Here we record some basic facts about exponential families.

The first fact says that for an arbitrary exponential family, the mean of the sufficient statistics is exactly the gradient of the log-partition function, and the covariance of the sufficient statistics is exactly the Hessian of the log-partition function.

Fact 33 (see, e.g., Wainwright and Jordan (2008)) *Let $X \sim P_\theta$ be an exponential family over $\mathcal{X}$ with sufficient statistics $T(x)$ and probability density function $P_\theta(x) = \exp\left(\langle T(x), \theta\rangle - A(\theta)\right)$, $\theta \in \mathbb{R}^d$. Let $\mu_T = \mathbf{E}[T(X)]$ and $\Sigma_T = \mathbf{Cov}[T(X)]$. Then, we have that $\nabla_\theta A(\theta) = \mu_T$ and $\nabla^2_\theta A(\theta) = \frac{\partial \mu_T}{\partial \theta} = \Sigma_T$.*

We include the proof for completeness.

Proof Let $Z(\theta) = \exp(A(\theta)) = \sum_{x\in\mathcal{X}} \exp\left(\langle T(x), \theta\rangle\right)$. From elementary calculation, we have that

$$\nabla A(\theta) = \nabla \ln Z(\theta) = \frac{\nabla Z(\theta)}{Z(\theta)} = \frac{\sum_{x\in\mathcal{X}} \exp(\langle T(x),\theta\rangle) T(x)}{Z(\theta)} = \mu_T\ ,$$

and

$$\nabla^2 A(\theta) = \frac{\partial \mu_T}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\frac{\sum_{x\in\mathcal{X}} \exp(\langle T(x),\theta\rangle)T(x)}{Z(\theta)}\right) = \frac{\sum_{x\in\mathcal{X}} \exp(\langle T(x),\theta\rangle)T(x)T(x)^T}{Z(\theta)} - \frac{\left(\sum_{x\in\mathcal{X}} \exp(\langle T(x),\theta\rangle)T(x)\right)\left(\sum_{x\in\mathcal{X}} \exp(\langle T(x),\theta\rangle)T(x)^T\right)}{Z(\theta)^2} = \mathbf{E}[T(X)T(X)^T] - \mu_T\mu_T^T = \Sigma_T \,.$$

■

The following fact connects the KL-divergence between two exponential families with their parameters in an explicit form.

Fact 34 (see, e.g., Wainwright and Jordan (2008)) *Let $P_\theta, P_{\theta'}$ be exponential families with probability density functions $P_\theta(x) = \exp\left(\langle T(x),\theta\rangle - A(\theta)\right)$ and $P_{\theta'}(x) = \exp\left(\langle T(x),\theta'\rangle - A(\theta')\right)$, where the parameters $\theta, \theta' \in \mathbb{R}^d$. Let $\mu_T = \mathbf{E}_{X\sim P_\theta}[T(X)]$, $\mu'_T = \mathbf{E}_{X\sim P_{\theta'}}[T(X)]$, $\Sigma_T = \mathbf{Cov}_{X\sim P_\theta}[T(X)]$, and $\Sigma'_T = \mathbf{Cov}_{X\sim P_{\theta'}}[T(X)]$. Then, we have that*

$$d_{\mathrm{KL}}(P_\theta, P_{\theta'}) = \langle \theta - \theta', \mu_T\rangle - A(\theta) + A(\theta').$$

Combining this with Fact 33, we obtain that $\nabla_{\theta'} d_{\mathrm{KL}}(P_\theta, P_{\theta'}) = \mu'_T - \mu_T$ and $\nabla^2_{\theta'} d_{\mathrm{KL}}(P_\theta, P_{\theta'}) = \Sigma'_T$.

Proof From the definition of KL-divergence, we have that

$$\begin{aligned} d_{\mathrm{KL}}(P_\theta, P_{\theta'}) &= \mathbf{E}_{X\sim P_\theta}\left[\ln\left(\frac{P_\theta(x)}{P_{\theta'}(x)}\right)\right] \\ &= \mathbf{E}_{X\sim P_\theta}[\langle T(X), \theta - \theta'\rangle] - A(\theta) + A(\theta') \\ &= \langle \theta - \theta', \mu_T\rangle - A(\theta) + A(\theta') \,. \end{aligned}$$

■

A.8. Proof of Lemma 10

From the definition of $L(\theta, \mu_T) = \langle \theta, \mu_T\rangle - A(\theta)$ and Fact 33, it follows that for any $\theta \in \Omega$, we have that

$$\nabla^2_\theta L(\theta, \mu_T) = -\nabla^2_\theta A(\theta) = -\mathbf{Cov}_{X\sim P_\theta}[T(X)] \preceq -c\, I \,.$$

Hence, for any fixed $\mu_T \in \mathbb{R}^d$, the objective function $L(\theta, \mu_T)$ is c-strongly concave and therefore has a unique maximizer $\theta_{\mu_T} \in \Omega$. From Fact 34, we have that

$$\begin{aligned} L(\theta^*, \mu^*_T) - L(\theta', \mu^*_T) &= \langle \theta^* - \theta', \mu^*_T\rangle - A(\theta^*) + A(\theta') = d_{\mathrm{KL}}(P_{\theta^*}, P_{\theta'}), \text{ and} \\ 0 \le L(\theta', \mu'_T) - L(\theta^*, \mu'_T) &= \langle \theta' - \theta^*, \mu'_T\rangle - A(\theta') + A(\theta^*) \,, \end{aligned}$$

where we used the fact that, given $\mu'_T \in \mathbb{R}^d$, $L(\theta, \mu'_T)$ attains its maximum at $\theta = \theta'$ over Ω. Adding the above two equations together, we get

$$\langle \theta^* - \theta', \mu^*_T - \mu'_T\rangle = \left(L(\theta^*, \mu^*_T) - L(\theta', \mu^*_T)\right) + \left(L(\theta', \mu'_T) - L(\theta^*, \mu'_T)\right) \ge d_{\mathrm{KL}}(P_{\theta^*}, P_{\theta'}) \,. \tag{2}$$

In addition, from Taylor's theorem, we can rewrite $d_{\mathrm{KL}}(P_{\theta^*}, P_{\theta'})$ as follows

$$\begin{aligned} d_{\mathrm{KL}}(P_{\theta^*}, P_{\theta'}) &= d_{\mathrm{KL}}(P_{\theta^*}, P_{\theta'}) - d_{\mathrm{KL}}(P_{\theta^*}, P_{\theta^*}) \\ &= \nabla_{\theta'} d_{\mathrm{KL}}(P_{\theta^*}, P_{\theta'})\Big|_{\theta'=\theta^*} + \frac{1}{2}(\theta' - \theta^*)^T \nabla^2_{\theta'} d_{\mathrm{KL}}(P_{\theta^*}, P_{\theta'})\Big|_{\theta'=\theta''}(\theta' - \theta^*) \\ &= \frac{1}{2}(\theta' - \theta^*)^T \mathbf{Cov}_{X\sim P_{\theta''}}[T(x)](\theta' - \theta^*) \\ &\geq \frac{c}{2}\|\theta' - \theta^*\|_2^2 \,, \end{aligned} \tag{3}$$

where $\theta'' = \lambda\theta' + (1-\lambda)\theta^*$ for some $0 \leq \lambda \leq 1$ and we apply Fact 34 in the third equality. Combining (2) and (3), we obtain that

$$\|\theta^* - \theta'\|_2 \, \|\mu_T^* - \mu_T'\|_2 \geq \langle \theta^* - \theta', \mu_T^* - \mu_T' \rangle \geq d_{\mathrm{KL}}(P_{\theta^*}, P_{\theta'}) \geq \frac{c}{2}\|\theta' - \theta^*\|_2^2 \,,$$

which implies that

$$\|\theta^* - \theta'\|_2 \leq \frac{2}{c}\|\mu_T^* - \mu_T'\|_2 \leq \frac{2\delta}{c} \,.$$

This completes the proof.

A.9. From Parameter Distance to Total Variation Distance

The following lemma shows that for any exponential family P_{θ^*}, if the sufficient statistics $T(x)$ is sub-exponential, then a good estimate for the parameter θ^* yields a good estimate in total variation distance.

Lemma 35 *Let P_{θ^*} be an exponential family over $\mathcal{X}$ with parameter $\theta^* \in \mathbb{R}^d$ and sufficient statistics $T(x)$. Let $\widehat{\theta} \in \mathbb{R}^d$ such that $\|\widehat{\theta} - \theta^*\|_2 \leq \delta$, for some sufficiently small constant $\delta > 0$. If for any unit vector $v \in \mathbb{R}^d$, $\mathbf{Pr}_{X\sim P_{\theta^*}}[|\langle v, T(X) - \mathbf{E}[T(X)]\rangle| > t] \leq 2\exp(-ct)$, for all $t > 0$, then $d_{\mathrm{TV}}(P_{\widehat{\theta}}, P_{\theta^*}) \leq c'\|\widehat{\theta} - \theta^*\|_2$, for some constant $c' > 0$.*

Proof Let $\theta = \widehat{\theta} - \theta^*$. Define $g(x) = \langle T(x), \theta\rangle - \mathbf{E}_{X\sim P_{\theta^*}}[\langle T(x), \theta\rangle]$. By definition, we have that $\mathbf{E}_{X\sim P_{\theta^*}}[g(X)] = 0$, and for any $x \in \mathcal{X}$,

$$\begin{aligned} \frac{P_{\widehat{\theta}}(x)}{P_{\theta^*}(x)} &= \frac{\exp(\langle T(x), \widehat{\theta}\rangle - A(\widehat{\theta}))}{\exp\left(\langle T(x), \theta^*\rangle - A(\theta^*)\right)} = \frac{\exp(\langle T(x), \theta\rangle)}{\exp(A(\widehat{\theta}) - A(\theta^*))} = \frac{\exp(\langle T(x), \theta\rangle)}{\sum_{x\in\mathcal{X}} \exp(\langle T(x), \widehat{\theta}\rangle - A(\theta^*))} \\ &= \frac{\exp(\langle T(x), \theta\rangle)}{\sum_{x\in\mathcal{X}} \exp(\langle T(x), \theta\rangle) \cdot \exp(\langle T(x), \theta^*\rangle - A(\theta^*))} = \frac{\exp(\langle T(x), \theta\rangle)}{\mathbf{E}_{X\sim P_{\theta^*}}[\exp(\langle T(X), \theta\rangle)]} \\ &= \frac{\exp(g(x))}{\mathbf{E}_{X\sim P_{\theta^*}}[\exp(g(X))]} = \exp(g(x))/w \,, \end{aligned}$$

where $w = \mathbf{E}_{X\sim P_{\theta^*}}[\exp(g(X))]$. In order to bound the total variation distance, we bound the χ^2-distance between $P_{\widehat{\theta}}$ and P_{θ^*}. Recall that for any two distributions p, q over $\mathcal{X}$, $\chi^2(p, q) \stackrel{\mathrm{def}}{=}$

$\int_{\mathcal{X}} \left(\frac{dp}{dq} - 1\right)^2 dq$, we have that

$$\chi^2(P_{\widehat{\theta}}, P_{\theta^*}) = \mathbf{E}_{X\sim P_{\theta^*}}\left[\left(\frac{P_{\widehat{\theta}}(X)}{P_{\theta^*}(X)} - 1\right)^2\right] = \frac{\mathbf{E}_{X\sim P_{\theta^*}}\left[(\exp(g(X)) - w)^2\right]}{w^2}$$
$$= \frac{\mathbf{E}_{X\sim P_{\theta^*}}[\exp(2g(X))]}{w^2} - 1 \leq \mathbf{E}_{X\sim P_{\theta^*}}[\exp(2g(X))] - 1 \,,$$

where we apply $w \geq 1$ in the last inequality, since

$$w = \mathbf{E}_{X\sim P_{\theta^1}}[\exp(g(X))] \geq \exp\left(\mathbf{E}_{X\sim P_{\theta^1}}[g(X)]\right) = 1$$

by Jensen's inequality. By our assumption, there is a constant $c_1 > 0$ such that

$$\mathbf{Pr}_{X\sim P_{\theta^*}}[|g(X) - \mathbf{E}[g(X)]| > t] \leq 2\exp\left(-\frac{c_1 t}{\|\theta\|_2}\right) \,.$$

Hence, from Fact 21, there is a constant $c_2 > 0$ such that as long as $|\lambda| \leq \frac{c_1}{c_2\|\theta\|_2}$, we will have that $\mathbf{E}_{X\sim P_{\theta^1}}[\exp(\lambda g(X))] \leq \exp(c_2^2\lambda^2\|\theta\|_2^2/c_1^2)$. Now we assume that $\|\theta\|_2^2 = \|\widehat{\theta} - \theta^*\|_2^2 \leq \delta^2 \leq c_1^2/4c_2^2$ and derive that

$$\chi^2(P_{\widehat{\theta}}, P_{\theta^*}) \leq \mathbf{E}_{X\sim P_{\theta^*}}[\exp(2g(X))] - 1 \leq \exp(4c_2^2\|\theta\|_2^2/c_1^2) - 1 \leq 8c_2^2\|\theta\|_2^2/c_1^2 \,,$$

where we apply the elementary inequality $e^x \leq 1 + 2x$, for $x \leq 1$. Therefore,

$$d_{\mathrm{TV}}(P_{\widehat{\theta}}, P_{\theta^*}) \leq \sqrt{\frac{\chi^2(P_{\widehat{\theta}}, P_{\theta^*})}{2}} \leq 2c_2\|\widehat{\theta} - \theta^*\|_2/c_1 \,.$$

■

Appendix B. Omitted Proofs from Section 3

B.1. Proof of Lemma 11

The following simple claim shows that under Condition 6, the likelihood function of the exponential family is smooth and strongly convex.

Claim 36 *Fix $\mu_T \in \mathbb{R}^d$. For any $\theta \in \Omega$, define $L(\theta, \mu_T) = \langle\theta, \mu_T\rangle - A(\theta)$, where $A(\theta)$ is the log-partition function for the exponential family P_θ with sufficient statistics $T(x)$. If Condition 6 holds, then $-L(\theta, \mu_T)$ is L-smooth and m-strongly convex, for some constants $L, m > 0$ independent of the vector μ_T.*

Proof Let $f(\theta) = -L(\theta, \mu_T)$ and we have that $\nabla f(\theta) = \mathbf{E}_{X\sim P_\theta}[T(X)] - \mu_T$ and $\nabla^2 f(\theta) = \mathbf{Cov}_{X\sim P_\theta}[T(X)]$. From the first statement in Condition 6, we know that $\mathbf{Cov}_{X\sim P_\theta}[T(X)] \succeq m\,I$ for some universal constant $m > 0$, and thus $f(\theta)$ is m-strongly convex. In addition, from the second statement in Condition 6, we know that there exists a constant $c > 0$ such that for any parameter $\theta \in \Omega$ and any unit vector $v \in \mathbb{S}^{d-1}$, $\mathbf{Pr}_{X\sim P_\theta}[|\langle v, T(X) - \mathbf{E}[T(X)]\rangle| > t] \leq$

$2\exp(-ct), \forall t > 0$. From Fact 21, we have that $\mathbf{Cov}_{X\sim P_\theta}[T(X)] \preceq L\, I$, for some universal constant $L > 0$ and thus $f(\theta)$ is L-smooth. ■

Since $-L(\theta, \mu_T)$ is L-smooth and m-strongly convex, one can apply Projected Gradient Descent (PGD) to efficiently compute the maximum likelihood estimator $\arg\max_{\theta\in\Omega} L(\theta, \mu_T)$ for any fixed $\mu_T \in \mathbb{R}^d$. A small wrinkle is that, in order to apply vanilla PGD (Algorithm 2), we need access to exact gradients and projections. In our setting, this is not possible in general: For general exponential families, it is computationally hard to compute $\nabla(-L(\theta, \mu_T)) = \mathbf{E}_{X\sim P_\theta}[T(X)] - \mu_T$ exactly. To address this minor issue, we need to slightly modify Algorithm 2 and its analysis, where we use sufficiently accurate approximations to the gradient and the projection.

Algorithm 3: Projected gradient descent for strongly convex smooth optimization with approximate gradient and projection

Input : L-smooth and m-strongly convex function f over Ω and parameters $\delta, \delta_1, \delta_2 > 0$.
Output: $\widehat{x} \in \Omega$ such that $\|\widehat{x} - x^*\|_2 \le \delta + \frac{\delta_2 + \delta_1/L}{1-\sqrt{1-m/L}}$, where $x^* = \arg\min_{x\in\Omega} f(x)$.

Let $x^0 \in \Omega$ be an arbitrary initial point and $T = O\left(\frac{L}{m}\log\left(\frac{\mathrm{diam}(\Omega)}{\delta}\right)\right)$.
for $t = 0$ *to* $T-1$ **do**
  Compute g^t such that $\|g^t - \nabla f(x^t)\|_2 \le \delta_1$.
  $r^t = x^t - \frac{1}{L}\, g^t$.
  Compute $x^{t+1} \in \Omega$ such that $\|x^{t+1} - P_\Omega(r^t)\|_2 \le \delta_2$, where $P_\Omega(r^t) = \arg\min_{x\in\Omega} \|x - r^t\|_2$.
end
return x^T;

The following simple claim adapts the analysis of PGD to work with approximate gradients and projections.

Claim 37 *Let $f : \Omega \to \mathbb{R}$ be L-smooth and m-strongly convex and $x^* = \arg\min_{x\in\Omega} f(x)$. The iterates in Algorithm 2 satisfy*

$$\|x^{t+1} - x^*\|_2 \le \delta_2 + \delta_1/L + \sqrt{1 - m/L}\|x^t - x^*\|_2 \,.$$

Therefore, after $T = O\left(\frac{L}{m}\log\left(\frac{\mathrm{diam}(\Omega)}{\delta}\right)\right)$ iterations, we have that $\|x^T - x^*\|_2 \le \delta + \frac{\delta_2+\delta_1/L}{1-\sqrt{1-m/L}}$.

Proof From Fact 24, we have that

$$\begin{aligned}
\left\|x^{t+1} - x^*\right\|_2 &\le \left\|x^{t+1} - P_\Omega(r^t)\right\|_2 + \left\|P_\Omega(r^t) - P_\Omega\left(x^t - \frac{1}{L}\nabla f(x^t)\right)\right\|_2 \\
&\quad + \left\|P_\Omega\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - x^*\right\|_2 \\
&\le \left\|x^{t+1} - P_\Omega(r^t)\right\|_2 + \left\|r^t - \left(x^t - \frac{1}{L}\nabla f(x^t)\right)\right\|_2 + \left\|P_\Omega\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - x^*\right\|_2 \\
&= \left\|x^{t+1} - P_\Omega(r^t)\right\|_2 + \frac{1}{L}\left\|g^t - \nabla f(x^t)\right\|_2 + \left\|P_\Omega\left(x^t - \frac{1}{L}\nabla f(x^t)\right) - x^*\right\|_2 \\
&\le \delta_2 + \delta_1/L + \sqrt{1 - m/L}\|x^t - x^*\|_2 \,,
\end{aligned}$$

where we apply $\|P_\Omega(x) - P_\Omega(y)\|_2 \leq \|x-y\|_2, \forall x, y \in \mathbb{R}^d$ in the second inequality. Therefore, we can write

$$\begin{aligned}\|x^T - x^*\|_2 - \frac{\delta_2 + \delta_1/L}{1-\sqrt{1-m/L}} &\leq \sqrt{1-m/L}\left(\|x^{T-1} - x^*\|_2 - \frac{\delta_2 + \delta_1/L}{1-\sqrt{1-m/L}}\right) \\ &\leq (1-m/L)^{T/2}\left(\|x^0 - x^*\|_2 - \frac{\delta_2 + \delta_1/L}{1-\sqrt{1-m/L}}\right) \\ &\leq (1-m/L)^{T/2}\mathrm{diam}(\Omega)\ .\end{aligned}$$

■

Claim 37 tells us that if we are able to efficiently approximate the projection of an arbitrary point in $\mathbb{R}^d$ to Ω and the gradient of the function, then we can efficiently solve the underlying minimization problem. From Condition 6, we can efficiently approximate the projection of any point in $\mathbb{R}^d$ within error $1/\mathrm{poly}(d)$. Note that for any fixed $\mu_T \in \mathbb{R}^d$, the gradient of the negative likelihood is equal to $\nabla(-L(\theta, \mu_T)) = \mathbf{E}_{X\sim P_\theta}[T(X)] - \mu_T$. Therefore, it suffices to show that for any given parameter θ, we can efficiently estimate the mean $\mathbf{E}_{X\sim P_\theta}[T(X)]$ within small error. This is done in the following claim:

Claim 38 *Let P, Q be distributions on $\mathbb{R}^d$. Assume that Q is sub-exponential and that $d_{\mathrm{TV}}(P,Q) \leq \gamma$ for some parameter $\gamma > 0$. Let μ and Σ denote the mean and covariance of distribution Q respectively. Let $X_1, \ldots, X_n$ be i.i.d. samples drawn from P and $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)(X_i - \widehat{\mu}_n)^T$ be the empirical mean and covariance. Then there exist constants $c_1, c_2 > 0$ such that the following holds:*

1. *With probability at least $1 - 2\exp(-t^2) - n\gamma$, we have that $\|\widehat{\mu}_n - \mu\|_2 \leq c_1 \max(\delta, \delta^2)$, where $\delta = \sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}$.*
2. *With probability at least $1 - 6\exp(-t) - n\gamma$, we have that*

$$\left\|\widehat{\Sigma}_n - \Sigma\right\|_2 \leq c_2 d\left(\sqrt{\frac{t+\log d}{n}} + \frac{((t+\log d)\log n)^2}{n}\right).$$

Proof Let $Y_1, \ldots, Y_n$ be n i.i.d. samples drawn from Q. Let

$$\mu_n = \frac{1}{n}\sum_{i=1}^n Y_i, \qquad \text{and} \qquad \Sigma_n = \frac{1}{n}\sum_{i=1}^n (Y_i - \mu_n)(Y_i - \mu_n)^T.$$

By the data processing inequality for the total variation distance, we can write

$$d_{\mathrm{TV}}(\mu_n, \widehat{\mu}_n) \leq d_{\mathrm{TV}}((X_1, \ldots, X_n), (Y_1, \ldots, Y_n)) \leq n d_{\mathrm{TV}}(P, Q) \leq n\gamma$$

and similarly

$$d_{\mathrm{TV}}(\Sigma_n, \widehat{\Sigma}_n) \leq d_{\mathrm{TV}}((X_1, \ldots, X_n), (Y_1, \ldots, Y_n)) \leq n d_{\mathrm{TV}}(P, Q) \leq n\gamma\ .$$

We pick optimal couplings $(\widehat{\mu}_n, \mu_n)$ and $(\widehat{\Sigma}_n, \Sigma_n)$. From Lemma 22, there exist constants $c_1, c_2 > 0$ such that

$$\begin{aligned}\mathbf{Pr}\left[\|\widehat{\mu}_n - \mu\|_2 > c_1 \max(\delta, \delta^2)\right] &\le \mathbf{Pr}\left[\widehat{\mu}_n \neq \mu_n\right] + \mathbf{Pr}\left[\|\mu_n - \mu\|_2 > c_1 \max(\delta, \delta^2)\right] \\ &\le 2\exp(-t^2) + n\gamma\,,\end{aligned}$$

and

$$\begin{aligned}&\mathbf{Pr}\left[\left\|\widehat{\Sigma}_n - \Sigma\right\|_2 > c_2 d\left(\sqrt{\frac{t+\log d}{n}} + \frac{((t+\log d)\log n)^2}{n}\right)\right] \\ &\le \mathbf{Pr}\left[\widehat{\Sigma}_n \neq \Sigma_n\right] + \mathbf{Pr}\left[\|\Sigma_n - \Sigma\|_2 > c_2 d\left(\sqrt{\frac{t+\log d}{n}} + \frac{((t+\log d)\log n)^2}{n}\right)\right] \\ &\le 6\exp(-t) + n\gamma\,.\end{aligned}$$

■

We are now ready to prove Lemma 11.

Proof [Proof of Lemma 11] Let $L(\theta, \mu'_T) = \langle \theta, \mu'_T\rangle - A(\theta), \forall \theta \in \Omega$ and $\theta' = \arg\max_{\theta\in\Omega} L(\theta, \mu'_T)$. By Lemma 10, we have that $\|\theta' - \theta^*\|_2 \le O(\delta)$. If we pick $\delta_1 = \delta_2 = \delta$ and apply Algorithm 3 to the function $-L(\theta, \mu'_T)$, it will return a point $\widehat{\theta} \in \Omega$ with $\|\widehat{\theta} - \theta'\|_2 \le O(\delta)$, since by Claim 36 $-L(\theta, \mu'_T)$ is L-smooth and m-strongly convex, for some universal constants $L, m > 0$. This implies that $\|\widehat{\theta} - \theta^*\|_2 \le \|\widehat{\theta} - \theta'\|_2 + \|\theta' - \theta^*\|_2 \le O(\delta)$.

Now we show that the above process is efficient and bound the failure probability. By Condition 6, $\mathrm{diam}(\Omega) \le \exp(d^c)$ for some constant $c > 0$. Given an arbitrary $\theta \in \Omega$, we can sample from a distribution within total variation distance $\gamma = \frac{\delta^2\zeta}{2d(d^2+\log(1/\delta))\log\left(\frac{4(d^c+\log(1/\delta))}{\zeta}\right)}$ from P_θ in time $\mathrm{poly}\left(\frac{d}{\delta\zeta}\right)$. Hence, if we pick $t = \sqrt{\log\left(\frac{4(d^c+\log(1/\delta))}{\zeta}\right)}$ and $n = \Omega(t^2 d/\delta^2)$ in Claim 38, we are able to estimate the gradient $\nabla_\theta(-L(\theta, \mu'_T)) = \mathbf{E}_{X\sim P_\theta}[T(X)] - \mu'_T$ within error δ with probability at least $1 - O\left(\frac{\zeta}{d^c+\log(1/\delta)}\right)$. Since there are $T = O\left(\frac{L}{m}\log\left(\frac{\mathrm{diam}(\Omega)}{\delta}\right)\right) = O(d^c + \log(1/\delta))$ iterations, by union bound, the algorithm will output a $\widehat{\theta}$ with $\|\widehat{\theta} - \theta^*\|_2 \le O(\delta)$ with probability at least $1 - \zeta$. ■

B.2. Proof of Proposition 16

Let $Z(\theta) = \exp(A(\theta))$ be the normalizing factor. Fix $i, j, k \in [d]$. We calculate the partial derivative $\frac{\partial(\Sigma_T)_{ij}}{\partial\theta_k}$ as follows.

$$\begin{aligned}
\frac{\partial(\Sigma_T)_{ij}}{\partial\theta_k} &= \frac{\partial}{\partial\theta_k}\left(\mathbf{E}_{X\sim P_\theta}[(T(X)-\mu_T)_i(T(X)-\mu_T)_j]\right) \\
&= \frac{\partial}{\partial\theta_k}\left(\frac{\sum_{x\in\mathcal{X}}\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i(T(x)-\mu_T)_j}{Z(\theta)}\right) \\
&= \sum_{x\in\mathcal{X}}\frac{\partial}{\partial\theta_k}\left(\frac{\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i(T(x)-\mu_T)_j}{Z(\theta)}\right) \\
&= \sum_{x\in\mathcal{X}}\frac{\frac{\partial}{\partial\theta_k}\left(\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i(T(x)-\mu_T)_j\right)}{Z(\theta)} \\
&\quad - \frac{\partial Z(\theta)}{\partial\theta_k}\sum_{x\in\mathcal{X}}\frac{\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i(T(x)-\mu_T)_j}{Z(\theta)^2}\,.
\end{aligned}$$

Noting that

$$\begin{aligned}
&\frac{\partial}{\partial\theta_k}\left(\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i(T(x)-\mu_T)_j\right) \\
&= \exp(\langle T(x),\theta\rangle)T(x)_k(T(x)-\mu_T)_i(T(x)-\mu_T)_j - \frac{\partial(\mu_T)_i}{\partial\theta_k}\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_j \\
&\quad - \frac{\partial(\mu_T)_j}{\partial\theta_k}\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i \\
&= \exp(\langle T(x),\theta\rangle)T(x)_k(T(x)-\mu_T)_i(T(x)-\mu_T)_j - (\Sigma_T)_{ik}\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_j \\
&\quad - (\Sigma_T)_{jk}\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i\,,
\end{aligned}$$

we have that

$$\begin{aligned}
&\sum_{x\in\mathcal{X}}\frac{\frac{\partial}{\partial\theta_k}\left(\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i(T(x)-\mu_T)_j\right)}{Z(\theta)} \\
&= \sum_{x\in\mathcal{X}}\frac{\exp(\langle T(x),\theta\rangle)T(x)_k(T(x)-\mu_T)_i(T(x)-\mu_T)_j}{Z(\theta)} - \frac{(\Sigma_T)_{ik}\sum_{x\in\mathcal{X}}\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_j}{Z(\theta)} \\
&\quad - \frac{(\Sigma_T)_{jk}\sum_{x\in\mathcal{X}}\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i}{Z(\theta)} \\
&= \mathbf{E}_{X\sim P_\theta}[T(x)_k(T(x)-\mu_T)_i(T(x)-\mu_T)_j] - (\Sigma_T)_{ik}\mathbf{E}_{X\sim P_\theta}[(T(x)-\mu_T)_j] \\
&\quad - (\Sigma_T)_{jk}\mathbf{E}_{X\sim P_\theta}[(T(x)-\mu_T)_i] \\
&= \mathbf{E}_{X\sim P_\theta}[T(x)_k(T(x)-\mu_T)_i(T(x)-\mu_T)_j]\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial(\Sigma_T)_{ij}}{\partial\theta_k} &= \sum_{x\in\mathcal{X}} \frac{\frac{\partial}{\partial\theta_k}\left(\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i(T(x)-\mu_T)_j\right)}{Z(\theta)} \\
&\quad - \frac{\partial Z(\theta)}{\partial\theta_k}\sum_{x\in\mathcal{X}} \frac{\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i(T(x)-\mu_T)_j}{Z(\theta)^2} \\
&= \mathbf{E}_{X\sim P_\theta}[T(x)_k(T(x)-\mu_T)_i(T(x)-\mu_T)_j] \\
&\quad - \left(\frac{\sum_{x\in\mathcal{X}}\exp(\langle T(x),\theta\rangle)T(x)_k}{Z(\theta)}\right)\left(\sum_{x\in\mathcal{X}}\frac{\exp(\langle T(x),\theta\rangle)(T(x)-\mu_T)_i(T(x)-\mu_T)_j}{Z(\theta)}\right) \\
&= \mathbf{E}_{X\sim P_\theta}[T(x)_k(T(x)-\mu_T)_i(T(x)-\mu_T)_j] - \mathbf{E}_{X\sim P_\theta}[T(x)_k]\mathbf{E}_{X\sim P_\theta}[(T(x)-\mu_T)_i(T(x)-\mu_T)_j] \\
&= \mathbf{E}_{X\sim P_\theta}[(T(x)-\mu_T)_i(T(x)-\mu_T)_j(T(x)-\mu_T)_k]\,.
\end{aligned}$$

This completes the proof.

B.3. Proof of Lemma 17

Let $\theta \in \Omega$ and P_θ be the corresponding exponential family with sufficient statistics $T(x)$. Let $\mu_T(\theta) = \mathbf{E}_{X\sim P_\theta}[T(x)]$ and $\Sigma_T(\theta) = \mathbf{Cov}_{X\sim P_\theta}[T(x)]$. Let $v \in \mathbb{S}^{d-1}$ be a unit vector such that $\|\Sigma_T(\theta^1) - \Sigma_T(\theta^2)\|_2 = \left|v^T(\Sigma_T(\theta^1) - \Sigma_T(\theta^2))v\right|$. Define $f(\theta) = v^T\Sigma_T(\theta)v$. By the mean value theorem, we have that

$$\begin{aligned}
\|\Sigma_T(\theta^1) - \Sigma_T(\theta^2)\|_2 &= \left|v^T\Sigma_T(\theta^1)v - v^T\Sigma_T(\theta^2)v\right| = |f(\theta^1) - f(\theta^2)| \\
&= |\langle\nabla f(\widetilde{\theta}), \theta^1 - \theta^2\rangle| \\
&\le \left\|\nabla f(\widetilde{\theta})\right\|_2 \cdot \|\theta^1 - \theta^2\|_2\,,
\end{aligned}$$

where $\widetilde{\theta} = \lambda\theta^1 + (1-\lambda)\theta^2$ for some $0 \le \lambda \le 1$. Therefore, we only need to show that $\left\|\nabla f(\widetilde{\theta})\right\|_2$ is upper bounded by a universal constant $c' > 0$. Let $w \in \mathbb{S}^{d-1}$ be the unit vector such that $\left\|\nabla f(\widetilde{\theta})\right\|_2 = \langle w, \nabla f(\widetilde{\theta})\rangle$. By our definition of function $f(\theta)$, we have that

$$\begin{aligned}
\left\|\nabla f(\widetilde{\theta})\right\|_2 = \langle w, \nabla f(\widetilde{\theta})\rangle &= \sum_{k=1}^{d}\frac{\partial f(\widetilde{\theta})}{\partial\theta_k}\cdot w_k = \sum_{k=1}^{d} v^T\left(\frac{\partial\Sigma_T(\widetilde{\theta})}{\partial\theta_k}\right)v\cdot w_k = \sum_{i,j,k\in[d]} v_iv_jw_k\frac{\partial(\Sigma_T)_{ij}(\widetilde{\theta})}{\partial\theta_k} \\
&= \sum_{i,j,k\in[d]} v_iv_jw_k\mathbf{E}_{X\sim P_{\widetilde{\theta}}}\left[(T(X)-\mu_T(\widetilde{\theta}))_i(T(X)-\mu_T(\widetilde{\theta}))_j(T(X)-\mu_T(\widetilde{\theta}))_k\right] \\
&= \mathbf{E}_{X\sim P_{\widetilde{\theta}}}[\langle T(X)-\mu_T(\widetilde{\theta}), v\rangle^2\langle T(X)-\mu_T(\widetilde{\theta}), w\rangle] \\
&\le \sqrt{\mathbf{E}_{X\sim P_{\widetilde{\theta}}}[\langle T(X)-\mu_T(\widetilde{\theta}), v\rangle^4]}\cdot\sqrt{\mathbf{E}_{X\sim P_{\widetilde{\theta}}}[\langle T(X)-\mu_T(\widetilde{\theta}), w\rangle^2]},
\end{aligned}$$

where we apply Proposition 16 in the fifth equality and the last inequality comes from Cauchy–Schwarz. From Fact 21, we know that both $\mathbf{E}_{X\sim P_{\widetilde{\theta}}}[\langle T(X)-\mu_T(\widetilde{\theta}), v\rangle^4]$ and $\mathbf{E}_{X\sim P_{\widetilde{\theta}}}[\langle T(X)-\mu_T(\widetilde{\theta}), w\rangle^2]$ are upper bounded by universal constants. Hence we obtain that $\left\|\nabla f(\widetilde{\theta})\right\|_2 \le c'$ for some universal constant $c' > 0$.

Appendix C. Omitted Proofs from Section 4

C.1. Proof of Lemma 18

Fact 39 (Götze et al. (2019)) *Let $X \sim P_\theta$ be an Ising model satisfying Dobrushin's condition and $\max_{i\in[d]} |\theta_i| \leq \alpha$, where $\alpha > 0$ is an absolute constant. Let $f : \{\pm 1\}^d \to \mathbb{R}$ be an arbitrary function. Define function $Df : \{\pm 1\}^d \to \mathbb{R}^d$ as $Df(x)_i = \frac{f(x_{i+})-f(x_{i-})}{2}, \forall x \in \{\pm 1\}^d, \forall i \in [d]$, where x_{i+} is the vector obtained from x by replacing the i-th coordinate with 1 and x_{i-} is the one that is obtained by replacing the i-th coordinate with -1. Define the function $Hf : \{\pm 1\}^d \to \mathbb{R}^{d\times d}$ as $Hf(x)_{ij} = D(Df(x)_j)(x)_i, \forall x \in \{\pm 1\}^d, \forall i, j \in [d]$. If $\mathbf{E}[\|Df(X)\|_2^2] \leq 1$ and $\|Hf(x)\|_F^2 \leq 1, \forall x \in \{\pm 1\}^d$, then there is a constant $c(\alpha, \eta) > 0$ such that*

$$\mathbf{Pr}[|f(X) - \mathbf{E}[f(X)]| > t] \leq 2\exp(-c(\alpha,\eta)\, t)\ ,$$

where $\eta > 0$ is the constant in Definition 3.

Proof [Proof of Lemma 18] Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix with zero diagonal and $b \in \mathbb{R}^d$ be such that $2\|A\|_F^2 + \|b\|_2^2 = 1$. Let $f(X) = (X-v)^T A(X-v) + b^T X$. From Fact 30, we can write

$$\begin{aligned}
\mathbf{E}\left[\|Df(X)\|_2^2\right] &= \frac{1}{4}\sum_{i=1}^d \mathbf{E}\left[(f(X_{i+}) - f(X_{i-}))^2\right] = \sum_{i=1}^d \mathbf{E}\left[\left(b_i + 2\sum_{j\neq i} A_{ij}(X_j - v_j)\right)^2\right] \\
&= \sum_{i=1}^d \mathbf{E}\left[\left(b_i + 2\sum_{j\neq i} A_{ij}(X_j - \mathbf{E}[X_j]) + 2\sum_{j\neq i} A_{ij}(\mathbf{E}[X_j] - v_j)\right)^2\right] \\
&\leq 3\sum_{i=1}^d b_i^2 + 12\sum_{i=1}^d \mathbf{Var}\left[\sum_{j\neq i} A_{ij}X_j\right] + 12\sum_{i=1}^d \left(\sum_{j\neq i} A_{ij}(\mathbf{E}[X_j] - v_j)\right)^2 \\
&\leq 3\sum_{i=1}^d b_i^2 + 12\sum_{i=1}^d \mathbf{Var}\left[\sum_{j\neq i} A_{ij}X_j\right] + 12\|A\|_F^2\|\mathbf{E}[X] - v\|_2^2 \\
&\leq 3\|b\|_2^2 + (c' + 12\delta^2)\|A\|_F^2\ ,
\end{aligned}$$

where in the first inequality we used the elementary identity $3(a^2+b^2+c^2) \geq (a+b+c)^2, \forall a, b, c \in \mathbb{R}$, and $c' > 0$ is an absolute constant. In addition, we have that

$$Hf(x)_{ij} = \frac{Df(x_{i+})_j - Df(x_{i-})_j}{2} = \frac{f(x_{i+,j+}) - f(x_{i+,j-}) - f(x_{i-,j+}) + f(x_{i-,j-})}{4} = A_{ij}\ ,$$

which implies that $\|Hf\|_F^2 = \sum_{i,j\in[d]} A_{ij}^2 = \|A\|_F^2$.

Hence, after a renormalization by $1/\sqrt{\max(3, c' + 12\delta^2)(\|A\|_F^2 + \|b\|_2^2)}$, the assumptions in Fact 39 are satisfied, and we have that

$$\mathbf{Pr}[|f(X) - \mathbf{E}[f(X)]| > t] \leq 2\exp\left(-\frac{ct}{\sqrt{\|A\|_F^2 + \|b\|_2^2}}\right)\ ,$$

where $c > 0$ is an absolute constant. ■

C.2. Proof of Theorem 19

By definition, we have that

$$\begin{aligned}\mathbf{Var}[(X-v)^T A(X-v)] &= \frac{1}{2}\mathbf{E}\left[\left((X-v)^T A(X-v) - (Y-v)^T A(Y-v)\right)^2\right] \\ &= \frac{1}{2}\mathbf{E}\left[\left((X-Y)^T A(X+Y-2v)\right)^2\right] , \end{aligned} \tag{4}$$

where Y is an independent copy of X. Let $S = \{i \in [d] \mid X_i = Y_i\}$. Then we can write

$$\begin{aligned}\mathbf{E}\left[\left((X-Y)^T A(X+Y-2v)\right)^2\right] &= 16\mathbf{E}\left[\mathbf{E}\left[\left(\sum_{i\notin S} X_i\left(\sum_{j\in S} A_{ij}(X_j - v_j) - \sum_{j\notin S} v_j A_{ij}\right)\right)^2\right]\Bigg| S\right] \\ &= 16\mathbf{E}\left[\mathbf{E}\left[\left(X_{-S}^T A^S W^S\right)^2 \mid S\right]\right] ,\end{aligned}$$

where $A^S_{ij} = A_{ij}$ for all $i \notin S, j \in [d]$ and $W^S_i = \mathbb{I}[i \in S]X_i - v_i$, for all $i \in [d]$.

Now for a fixed subset $S \subseteq [d]$, we calculate the conditional probability $\mathbf{Pr}[X = x \mid S]$. By our definition of S, we have that

$$\begin{aligned}\mathbf{Pr}[X = x \mid S] &= \mathbf{Pr}[X = x \wedge Y_S = x_S \wedge Y_{-S} = -x_{-S} \mid S] \\ &= \frac{\mathbf{Pr}[[X = x \wedge Y_S = x_S \wedge Y_{-S} = -x_{-S}]}{\mathbf{Pr}[S]} \\ &= \frac{\exp\left(\sum_{i\in S, j\in S}\theta_{ij}x_i x_j + \sum_{i\notin S, j\notin S}\theta_{ij}x_i x_j + 2\sum_{i\in S}\theta_i x_i\right)}{Z(\theta)^2\mathbf{Pr}[S]} ,\end{aligned}$$

where $Z(\theta)$ is the partition function of Ising model P_θ. Therefore conditioning on S, the marginal distribution of X is exactly an Ising model distribution with parameters

$$\theta^S_{ij} = \begin{cases} 2\theta_{ij} & i \in S, j \in S \text{ or } i \notin S, j \notin S, \\ 0 & \text{otherwise} , \end{cases} \qquad \text{and} \qquad \theta^S_i = \begin{cases} 2\theta_i & i \in S, \\ 0 & i \notin S, \end{cases}$$

which implies that conditioning on S, the marginal distribution X_S and X_{-S} are independent $(2M, 2\alpha)$-bounded Ising model distributions and $\mathbf{E}[X_{-S} \mid S] = 0$. Therefore, from Fact 32, there is a universal constant $c_1(M, \alpha) > 0$ such that

$$\begin{aligned}\mathbf{E}\left[\left(X_{-S}^T A^S W^S\right)^2 \mid S\right] &= \mathbf{E}_{X_S}\left[\mathbf{E}_{X_{-S}}\left[\left(X_{-S}^T A^S W^S\right)^2 \mid X_S, S\right] \mid S\right] \\ &\geq \mathbf{E}_{X_S}\left[\lambda_{\min}\left(\mathbf{E}_{X_{-S}}\left[X_{-S}X_{-S}^T \mid S\right]\right)\|A^S W^S\|_2^2 \mid S\right] \\ &\geq c_1(M,\alpha)\mathbf{E}\left[\|A^S W^S\|_2^2 \mid S\right] ,\end{aligned}$$

where $\lambda_{\min}\left(\mathbf{E}_{X_{-S}}\left[X_{-S}X_{-S}^T \mid S\right]\right)$ denotes the minimum eigenvalue of $\mathbf{E}_{X_{-S}}\left[X_{-S}X_{-S}^T \mid S\right]$ and in the first inequality, we use the fact that conditioning on S, X_S and X_{-S} are independent.

Therefore, we have that

$$\mathbf{E}\left[\left((X-Y)^T A(X+Y-2v)\right)^2\right] = 16\mathbf{E}\left[\mathbf{E}\left[\left(X_{-S}^T A^S W^S\right)^2 \mid S\right]\right] \tag{5}$$

$$\geq 16c_1(M,\alpha)\mathbf{E}\left[\mathbf{E}\left[\|A^S W^S\|_2^2 \mid S\right]\right] = 16c_1(M,\alpha)\mathbf{E}\left[\mathbf{E}\left[\sum_{i\notin S}\left(\sum_{j\in[d]} a_{ij}(\mathbb{I}[j\in S]X_j - v_j)\right)^2 \middle| S\right]\right]$$

$$= 4c_1(M,\alpha)\mathbf{E}\left[\|A'(X+Y-2v)\|_2^2\right] , \tag{6}$$

where $A'_{ij} = \mathbb{I}[X_i \neq Y_i]A_{ij}$ for all $i,j \in [d]$. Now we write $A' = [\mathbb{I}[X_1 \neq Y_1]a^1, \cdots, \mathbb{I}[X_d \neq Y_d]a^d]^T$, where $(a^i)^T$ denotes the i-th row vector of matrix A. By linearity of expectation, we have that

$$\mathbf{E}\left[\|A'(X+Y-2v)\|_2^2\right] = \sum_{i=1}^d \mathbf{E}\left[\langle \mathbb{I}[X_i \neq Y_i]a^i, X+Y-2v\rangle^2\right] . \tag{7}$$

By the law of total expectation, we can write

$$\mathbf{E}\left[\langle \mathbb{I}[X_i \neq Y_i]a^i, X+Y-2v\rangle^2\right] \tag{8}$$

$$= \mathbf{Pr}[X_i = 1, Y_i = -1]\cdot \mathbf{E}\left[\langle a^i, X+Y-2v\rangle^2 \mid X_i = 1, Y_i = -1\right]$$
$$+ \mathbf{Pr}[X_i = -1, Y_i = 1]\cdot \mathbf{E}\left[\langle a^i, X+Y-2v\rangle^2 \mid X_i = -1, Y_i = 1\right]$$

$$\geq 2\left(\frac{\exp(-2(\alpha+M))}{1+\exp(-2(\alpha+M))}\right)^2 \mathbf{E}\left[\langle a^i, X+Y-2v\rangle^2 \mid X_i = 1, Y_i = -1\right] , \tag{9}$$

where the inequality comes from Proposition 26 and the fact that Y is an independent copy of X.

Now we try to bound $\mathbf{E}\left[\langle a^i, X+Y-2v\rangle^2 \mid X_i = 1, Y_i = -1\right]$. From Fact 25, conditioning on $X_i = q \in \{\pm 1\}$, X_{-i} is an Ising model with parameter θ' satisfying the following property

$$\max_{j\in[d]\setminus\{i\}} |\theta'_j| \leq M+\alpha, \qquad \max_{j\in[d]\setminus\{i\}} \sum_{k\in[d]\setminus\{i,j\}} |\theta'_{jk}| \leq M ,$$

which implies that conditioning on $X_i = q$, X_{-i} is an $(M, M+\alpha)$-bounded Ising model. Note that X_{-i} and Y_{-i} are independent, conditioning on $X_i = -1, Y_i = 1$, from Fact 32, there is a constant $c_2(M,\alpha) > 0$ such that

$$\begin{aligned}\mathbf{E}\left[\langle a^i, X+Y-2v\rangle^2 \mid X_i = 1, Y_i = -1\right] &\geq \mathbf{Var}\left[\langle a^i, X+Y\rangle \mid X_i = 1, Y_i = -1\right] \\ &= \mathbf{Var}[\langle a^i, X\rangle \mid X_i = 1] + \mathbf{Var}[\langle a^i, Y\rangle \mid Y_i = -1] \\ &\geq c_2(M,\alpha)\|a^i\|_2^2 ,\end{aligned} \tag{10}$$

Combine (7), (8) and (10), we obtain that

$$\begin{aligned}\mathbf{E}\left[\|A'(X+Y-2v)\|_2^2\right] &= \sum_{i=1}^d \mathbf{E}\left[\langle \mathbb{I}[X_i \neq Y_i]a^i, X+Y-2v\rangle^2\right] \\ &\geq 2\left(\frac{\exp(-2(\alpha+M))}{1+\exp(-2(\alpha+M))}\right)^2 \sum_{i=1}^d \mathbf{E}\left[\langle a^i, X+Y-2v\rangle^2 \mid X_i = 1, Y_i = -1\right] \\ &\geq 2\left(\frac{\exp(-2(\alpha+M))}{1+\exp(-2(\alpha+M))}\right)^2 c_2(M,\alpha)\|A\|_F^2 .\end{aligned} \tag{11}$$

Combine (4), (5) and (11), we know that there exists a constant $c(M,\alpha) > 0$ such that

$$\mathbf{Var}[(X-v)^T A(X-v)] = \frac{1}{2}\mathbf{E}\left[\left((X-Y)^T A(X+Y-2v)\right)^2\right] \geq c(M,\alpha)\|A\|_F^2 .$$

C.3. Robustly Learning Ising Models with Non-zero External Field

In this section, we provide an efficient algorithm that robustly learns an Ising model with nonzero-external field. The main theorem of this section is the following:

Theorem 40 *Let P_{θ^*} be an Ising model. Let $\alpha > 0$ and $0 < M < 1$ be universal constants such that $\max_{i\in[d]} \sum_{j\neq i} |\theta^*_{ij}| \leq M$ and $\max_{i\in[d]} |\theta^*_i| \leq \alpha$. Let $0 < \epsilon \leq \epsilon_0$ for some universal constant ϵ_0 and S' be an ϵ-corrupted set of samples from P_{θ^*}. Let N be the size of S'. If there is a constant $c_0 > 0$ such that*

$$4\left(\frac{M}{1-M} + O(\sqrt{\epsilon})\right)^2 \leq (1-c_0)\left(8\left(\frac{\exp(-2(\alpha+2M))}{1+\exp(-2(\alpha+2M))}\right)^2 - \frac{2M}{1-M} - c_0\right), \quad (12)$$

then there is a $\mathrm{poly}(d/\epsilon)$ time algorithm that, for some $N = \widetilde{O}(d^2/\epsilon^2)$[2], on input S' and ϵ, returns an Ising model $P_{\widehat{\theta}}$ such that with probability at least $99/100$, we have that $d_{\mathrm{TV}}(P_{\widehat{\theta}}, P_{\theta^*}) \leq O(\epsilon\log(1/\epsilon))$. In addition, $P_{\widehat{\theta}}$ satisfies the Dobrushin's condition.

Intuitively, the theorem states that as long as the dependencies among each point and the external fields are sufficiently small, we can properly learn the Ising model distribution within small total variation distance in the strong contamination model. However, due to technical reasons, the constraint (12) is much stronger than the Dobrushin's condition and we are not able to obtain an efficient algorithm that learns the parameter θ^* here.

Similar to the zero-external field case, we view the Ising model distribution P_θ as an instance of an exponential family and try to apply Algorithm 1. However, if we choose the sufficient statistics $T(x) = ((x_i x_j)_{1\leq i<j\leq d}, (x_i)_{i\in[d]})$ in the straightforward way, the first statement in Condition 6 will not hold. For instance, consider the Ising model P_θ with $\theta_{ij} = 0, \forall i,j \in [d]$ and $\theta_i = \beta, \forall i \in [d]$ for some $\beta > 0$, such that $\mathbf{E}_{X\sim P_\theta}[X_i] = 1/2, \forall i \in [d]$. Let $A \in \mathbb{R}^{d\times d}$ be such that $A_{ij} = \frac{1}{\sqrt{d(d-1)(d+1)}}, \forall i \neq j$ and $b_i = -\sqrt{\frac{d-1}{d(d+1)}}, \forall i \in [d]$. In this case, we have that $2\|A\|_F^2 + \|b\|_2^2 = 1$ and

$$\begin{aligned}\mathbf{Var}[X^T AX + b^T X] &= \mathbf{Var}[(X-v)^T A(X-v) + (2Av+b)^T X] \\ &= \mathbf{Var}[(X-v)^T A(X-v)] \\ &\leq c\|A\|_F^2 = \frac{c}{d+1} ,\end{aligned}$$

where the last inequality comes from Fact 30 and $c > 0$ is an absolute constant.

2. Here we fix M, α and c_0 to be universal constants. Therefore we will suppress any possible dependence on M, α and c_0 in our asymptotic notation in this section.

To address this issue, we rewrite the density of an Ising model as the following "v-centered form". Let $v \in \mathbb{R}^d$ be an arbitrary fixed vector. By definition of the Ising model, we have that

$$\begin{aligned} P_\theta(x) &= \frac{1}{Z(\theta)} \exp\left(\frac{1}{2}\sum_{i,j\in[d]} \theta_{ij}x_i x_j + \sum_{i=1}^d \theta_i x_i\right) \\ &= \frac{1}{Z(\theta)} \exp\left(\frac{1}{2}\sum_{i,j\in[d]} \theta_{ij}(x_i - v_i)(x_j - v_j) + \sum_{i=1}^d \left(\theta_i + \sum_{j\in[d]} \theta_{ij}v_j\right)x_i\right) \\ &= \frac{1}{Z(\theta)} \exp\left(\frac{1}{2}(x-v)^T J(\theta)(x-v) + h(\theta)^T x\right) , \end{aligned}$$

where $J(\theta)_{ij} = \theta_{ij}, \forall i,j \in [d]$ and $h(\theta)_i = \theta_i + \sum_{j\in[d]} \theta_{ij}v_j$. If we write the probability density function $P_\theta(x)$ in the "v-centered form" as an instance of an exponential family, the sufficient statistics $T(x)$ will be

$$T(x) = ((x_i - v_i)(x_j - v_j)_{1\le i<j\le d}, (x_i)_{1\le i\le d}) ,$$

and the projection of $T(x)$ on a fixed direction is

$$(X-v)^T A(X-v) + b^T X ,$$

where $A \in \mathbb{R}^{d\times d}$ is a symmetric matrix with zero diagonal and $b \in \mathbb{R}^d$ with $2\|A\|_F^2 + \|b\|_2^2 = 1$.

In this way, by taking v to be an estimate of $\mathbf{E}_{X\sim P_{\theta^*}}[X]$, we are able to prove the following lower bound for covariance of the sufficient statistics $T(x)$ and then apply Algorithm 1 to robustly learn the parameter $J(\theta^*)$ and $h(\theta^*)$ in the "v-centered form".

Theorem 41 *Let $X \sim P_\theta$ be an Ising model. Let $\alpha > 0$ and $0 < M < 1$ be absolute constants such that $\max_{i\in[d]} \sum_{j\neq i} |\theta_{ij}| \le M$ and $\max_{i\in[d]} |\theta_i| \le \alpha$. Let $v \in \mathbb{R}^d$ be a vector such that $\|v - \mathbf{E}[X]\|_2 \le \delta$ for some constant $\delta > 0$. If there is a constant $c_0 > 0$ such that*

$$4\left(\frac{M}{1-M} + \delta\right)^2 \le (1-c_0)\left(8\left(\frac{\exp(-2(\alpha+2M))}{1+\exp(-2(\alpha+2M))}\right)^2 - \frac{2M}{1-M} - c_0\right) ,$$

then there exists another constant $c(\alpha, M, c_0) > 0$ such that

$$\mathbf{Var}[(X-v)^T A(X-v) + b^T X] \ge c(\alpha, M, c_0)(\|A\|_F^2 + \|b\|_2^2)$$

holds for all symmetric $A \in \mathbb{R}^{d\times d}$ with zero diagonal and $b \in \mathbb{R}^d$.

Proof By definition, we have that

$$\begin{aligned} \mathbf{Var}[(X-v)^T A(X-v) + b^T X] &= \frac{1}{2}\mathbf{E}\left[\left((X-v)^T A(X-v) - (Y-v)^T A(Y-v) + b^T X - b^T Y\right)^2\right] \\ &= \frac{1}{2}\mathbf{E}\left[\left((X-Y)^T\left(A(X+Y-2v)+b\right)\right)^2\right] , \end{aligned} \tag{13}$$

where Y is an independent copy of X. Let $S = \{i \in [d] \mid X_i = Y_i\}$ and we can write

$$\begin{aligned}
&\mathbf{E}\left[\left((X-Y)^T\left(A(X+Y-2v)+b\right)\right)^2\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\left(\sum_{i\notin S} X_i\left(b_i + \sum_{j\in S} 2a_{ij}(X_j - v_j) - \sum_{j\notin S} 2v_j a_{ij}\right)\right)^2\right]\middle| S\right] \\
&= 4\mathbf{E}\left[\mathbf{E}\left[\left(X_{-S}^T\left(A^S W^S + b_{-S}\right)\right)^2 \mid S\right]\right],
\end{aligned}$$

where $A^S_{ij} = A_{ij}$ for all $i \notin S, j \in [d]$ and $W^S_i = 2\left(\mathbb{I}[i \in S]X_i - v_i\right)$ for all $i \in [d]$.

Now for a fixed subset $S \subseteq [d]$, we calculate the conditional probability $\mathbf{Pr}[X = x \mid S]$. By definition of S, we have that

$$\begin{aligned}
\mathbf{Pr}[X = x \mid S] &= \mathbf{Pr}[X = x \wedge Y_S = x_S \wedge Y_{-S} = -x_{-S} \mid S] \\
&= \frac{\mathbf{Pr}[[X = x \wedge Y_S = x_S \wedge Y_{-S} = -x_{-S}]}{\mathbf{Pr}[S]} \\
&= \frac{\exp\left(2\sum_{i\in S, j\in S}\theta_{ij}x_i x_j + 2\sum_{i\notin S, j\notin S}\theta_{ij}x_i x_j + 2\sum_{i\in S}\theta_i x_i\right)}{Z(\theta)^2\mathbf{Pr}[S]},
\end{aligned}$$

where $Z(\theta)$ is the partition function of Ising model P_θ. Therefore conditioning on S, the marginal distribution of X is exactly an Ising model distribution with parameters

$$\theta^S_{ij} = \begin{cases} 2\theta_{ij} & i \in S, j \in S \text{ or } i \notin S, j \notin S, \\ 0 & \text{otherwise}, \end{cases} \qquad \text{and} \qquad \theta^S_i = \begin{cases} 2\theta_i & i \in S, \\ 0 & i \notin S, \end{cases}$$

which implies that conditioning on S, the marginal distribution X_S and X_{-S} are independent $(2M, 2\alpha)$-bounded Ising model distributions and $\mathbf{E}[X_{-S} \mid S] = 0$. Therefore, from Fact 32, there is a constant $c_1(M, \alpha) > 0$ such that

$$\begin{aligned}
\mathbf{E}\left[\left(X_{-S}^T\left(A^S W^S + b_{-S}\right)\right)^2 \mid S\right] &= \mathbf{E}_{X_S}\left[\mathbf{E}_{X_{-S}}\left[\left(X_{-S}^T\left(A^S W^S + b_{-S}\right)\right)^2 \mid X_S, S\right] \mid S\right] \\
&\geq \mathbf{E}_{X_S}\left[\lambda_{\min}\left(\mathbf{E}_{X_{-S}}\left[X_{-S}X_{-S}^T \mid S\right]\right)\|A^S W^S + b_{-S}\|_2^2 \mid S\right] \\
&\geq c_1(M,\alpha)\mathbf{E}\left[\|A^S W^S + b_{-S}\|_2^2 \mid S\right],
\end{aligned}$$

where $\lambda_{\min}\left(\mathbf{E}_{X_{-S}}\left[X_{-S}X_{-S}^T \mid S\right]\right)$ denotes the minimum eigenvalue of $\mathbf{E}_{X_{-S}}\left[X_{-S}X_{-S}^T \mid S\right]$ and in the first inequality, we use the fact that conditioning on S, X_S and X_{-S} are independent. Therefore, we have that

$$\begin{aligned}
&\mathbf{E}\left[\left((X-Y)^T\left(A(X+Y-2v)+b\right)\right)^2\right] && (14)\\
&= 4\mathbf{E}\left[\mathbf{E}\left[\left(X_{-S}^T\left(A^S W^S + b_{-S}\right)\right)^2 \mid S\right]\right] \\
&\geq 4c_1(M,\alpha)\mathbf{E}\left[\mathbf{E}\left[\|A^S W^S + b_{-S}\|_2^2 \mid S\right]\right] && (15)\\
&= c_1(M,\alpha)\mathbf{E}\left[\mathbf{E}\left[\sum_{i\notin S}\left(b_i + \sum_{j\in[d]} 2A_{ij}(\mathbb{I}[j\in S]X_j - v_j)\right)^2\middle| S\right]\right] \\
&= c_1(M,\alpha)\mathbf{E}\left[\|A'(X+Y-2v)+b'\|_2^2\right], && (16)
\end{aligned}$$

where $A'_{ij} = \mathbb{I}[X_i \neq Y_i]A_{ij}$ for all $i, j \in [d]$ and $b'_i = \mathbb{I}[X_i \neq Y_i]b_i$ for all $i \in [d]$. Now we write $A' = [\mathbb{I}[X_1 \neq Y_1]a^1, \cdots, \mathbb{I}[X_d \neq Y_d]a^d]^T$, where $(a^i)^T$ denotes the i-th row vector of matrix A. By linearity of expectation, we have that

$$\mathbf{E}\left[\|A'(X+Y-2v)+b'\|_2^2\right] = \sum_{i=1}^{d} \mathbf{E}\left[\left(\langle \mathbb{I}[X_i \neq Y_i]a^i, X+Y-2v\rangle + b'_i\right)^2\right]. \tag{17}$$

Fix some $i \in [d]$. Note that Y is an independent copy of X, we can write

$$\begin{aligned}
&\mathbf{E}\left[\left(\langle \mathbb{I}[X_i \neq Y_i]a^i, X+Y-2v\rangle + b'_i\right)^2\right] &(18)\\
&= \mathbf{Pr}[X_i = 1, Y_i = -1] \cdot \mathbf{E}\left[\left(\langle a^i, X+Y-2v\rangle + b_i\right)^2 \mid X_i = 1, Y_i = -1\right]\\
&\quad + \mathbf{Pr}[X_i = -1, Y_i = 1] \cdot \mathbf{E}\left[\left(\langle a^i, X+Y-2v\rangle + b_i\right)^2 \mid X_i = -1, Y_i = 1\right]\\
&\geq 2\left(\frac{\exp(-2(\alpha+M))}{1+\exp(-2(\alpha+M))}\right)^2 \cdot \mathbf{E}\left[\left(\langle a^i, X+Y-2v\rangle + b_i\right)^2 \mid X_i = -1, Y_i = 1\right], &(19)
\end{aligned}$$

where the inequality comes from Proposition 26.

Now we bound $\mathbf{E}\left[\left(\langle a^i, X+Y-2v\rangle + b_i\right)^2 \mid X_i = -1, Y_i = 1\right]$ as follows. From Proposition 25, we know that conditioning on $X_i = q \in \{\pm 1\}$, X_{-i} is an Ising model over $\{\pm 1\}^{d-1}$ with parameter θ' satisfying the following property

$$\max_{j \in [d]\setminus\{i\}} |\theta'_j| \leq M + \alpha, \qquad \max_{j \in [d]\setminus\{i\}} \sum_{k \in [d]\setminus\{i,j\}} |\theta'_{jk}| \leq M,$$

which implies that conditioning on $X_i = q$, X_{-i} is an $(M, M+\alpha)$-bounded Ising model. Let μ^1_{-i} denote the conditional expectation over x_{-i} conditioning on $x_i = 1$ and μ^{-1}_{-i} denote the conditional expectation over x_{-i} conditioning on $x_i = -1$. Note that X_{-i} and Y_{-i} are independent conditioning on $X_i = -1, Y_i = 1$, we have that

$$\begin{aligned}
&\mathbf{E}\left[\left(\langle a^i, X+Y-2v\rangle + b_i\right)^2 \mid X_i = 1, Y_i = -1\right]\\
&= \mathbf{Var}\left[\langle a^i, X+Y-2v\rangle + b_i \mid X_i = 1, Y_i = -1\right] + \mathbf{E}\left[\langle a^i, X+Y-2v\rangle + b_i \mid X_i = 1, Y_i = -1\right]^2\\
&= \mathbf{Var}[\langle a^i_{-i}, X_{-i}\rangle \mid X_i = 1] + \mathbf{Var}[\langle a^i_{-i}, Y_{-i}\rangle \mid Y_i = -1] + \left(b_i + \langle a^i_{-i}, \mu^1_{-i} + \mu^{-1}_{-i} - 2v_{-i}\rangle\right)^2\\
&\geq \mathbf{Var}[\langle a^i_{-i}, X_{-i}\rangle \mid X_i = 1] + \mathbf{Var}[\langle a^i_{-i}, Y_{-i}\rangle \mid Y_i = -1] + b_i^2 + 2b_i\langle a^i_{-i}, \mu^1_{-i} + \mu^{-1}_{-i} - 2v_{-i}\rangle\\
&\geq \mathbf{Var}[\langle a^i_{-i}, X_{-i}\rangle \mid X_i = 1] + \mathbf{Var}[\langle a^i_{-i}, Y_{-i}\rangle \mid Y_i = -1] + b_i^2 - 2|b_i|\|a^i\|_2\|\mu^1_{-i} + \mu^{-1}_{-i} - 2v_{-i}\|_2,
\end{aligned}$$

where we use $A_{ii} = 0, \forall i \in [d]$.

Let $\mu = \mathbf{E}[X]$ and thus $\|\mu - v\|_2 \leq \delta$ by our assumption. From Fact 31, we know that

$$\begin{aligned}
\|\mu^1_{-i} + \mu^{-1}_{-i} - 2v_{-i}\|_2 &\leq \|\mu^1_{-i} + \mu^{-1}_{-i} - 2\mu_{-i}\|_2 + 2\|\mu_{-i} - v_{-i}\|_2\\
&= (1 - \mathbf{Pr}[X_i = 1])\|\mu^1_{-i} - \mu^{-1}_{-i}\|_2 + 2\|\mu_{-i} - v_{-i}\|_2\\
&\leq \|\mu^1_{-i} - \mu^{-1}_{-i}\|_1 + 2\delta\\
&\leq \frac{2M}{1-M} + 2\delta.
\end{aligned}$$

From Fact 31 and Proposition 25, we have that

$$\begin{aligned}
&\mathbf{Var}[\langle a^i_{-i}, X_{-i}\rangle \mid X_i = q] = \sum_{j\neq i,k\neq i} A_{ij}A_{ik}\mathbf{Cov}(X_j, X_k \mid X_i = q)\\
&\geq \sum_{j\neq i} A_{ij}^2\mathbf{Var}[X_j \mid X_i = q] - \sum_{j\neq i,k\neq i,k\neq j} |A_{ij}||A_{ik}||\mathbf{Cov}(X_j, X_k \mid X_i = q)|\\
&\geq \sum_{j\neq i} A_{ij}^2\mathbf{Var}[X_j \mid X_i = q] - \sum_{j\neq i,k\neq i,k\neq j} \frac{(A_{ij}^2 + A_{ik}^2)|\mathbf{Cov}(X_j, X_k \mid X_i = q)|}{2}\\
&= \sum_{j\neq i} A_{ij}^2\left(\mathbf{Var}[X_j \mid X_i = q] - \sum_{k\neq j,k\neq i} |\mathbf{Cov}(X_j, X_k \mid X_i = q)|\right)\\
&\geq \sum_{j\neq i} A_{ij}^2\left(\mathbf{Var}[X_j \mid X_i = q] - \frac{M}{1-M}\right)\\
&\geq \left(4\left(\frac{\exp(-2(\alpha+2M))}{1+\exp(-2(\alpha+2M))}\right)^2 - \frac{M}{1-M}\right)\|a^i\|_2^2.
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\mathbf{E}\left[\left(\langle a^i, X+Y-2v\rangle + b_i\right)^2 \mid X_i = 1, Y_i = -1\right]\\
&\geq \mathbf{Var}[\langle a^i_{-i}, X_{-i}\rangle \mid X_i = 1] + \mathbf{Var}[\langle a^i_{-i}, Y_{-i}\rangle \mid Y_i = -1] + b_i^2 - 2|b_i|\|a^i\|_2\|\mu^1_{-i} + \mu^{-1}_{-i} - 2v_{-i}\|_2\\
&\geq \left(8\left(\frac{\exp(-2(\alpha+2M))}{1+\exp(-2(\alpha+2M))}\right)^2 - \frac{2M}{1-M}\right)\|a^i\|_2^2 + b_i^2 - 2|b_i|\|a^i\|_2\left(\frac{2M}{1-M} + 2\delta\right)\\
&\geq c_0(\|a^i\|_2^2 + b_i^2)\,, \qquad (20)
\end{aligned}$$

as long as

$$4\left(\frac{M}{1-M} + \delta\right)^2 \leq (1-c_0)\left(8\left(\frac{\exp(-2(\alpha+2M))}{1+\exp(-2(\alpha+2M))}\right)^2 - \frac{2M}{1-M} - c_0\right).$$

Combine (17), (18) and (20), we obtain that

$$\begin{aligned}
&\mathbf{E}\left[\|A'(X+Y-2v) + b'\|_2^2\right] = \sum_{i=1}^d \mathbf{E}\left[\left(\langle \mathbb{I}[X_i \neq Y_i]\alpha^i, X+Y-2v\rangle + b'_i\right)^2\right]\\
&\geq 2\left(\frac{\exp(-2(\alpha+M))}{1+\exp(-2(\alpha+M))}\right)^2 \sum_{i=1}^d \mathbf{E}\left[\left(\langle a^i, X+Y-2v\rangle + b_i\right)^2 \mid X_i = -1, Y_i = 1\right]\\
&\geq 2c_0\left(\frac{\exp(-2(\alpha+M))}{1+\exp(-2(\alpha+M))}\right)^2 (\|A\|_F^2 + \|b\|_2^2)\,. \qquad (21)
\end{aligned}$$

Combine (13), (14) and (21), we know that there is a constant $c(\alpha, M, c_0)$ such that

$$\begin{aligned}
\mathbf{Var}[(X-v)^T A(X-v) + b^T X] &= \frac{1}{2}\mathbf{E}\left[\left((X-Y)^T\left(A(X+Y-2v)+b\right)\right)^2\right]\\
&\geq c(\alpha, M, c_0)(\|A\|_F^2 + \|b\|_2^2).
\end{aligned}$$

■

Proof [Proof of Theorem 40] From Fact 29, we know that $X \sim P_{\theta^*}$ is sub-Gaussian, and thus $\mathbf{Cov}_{X\sim P_{\theta^*}} \preceq c_0\, I$, for some universal constant $c_0 > 0$. Hence, we can apply the robust mean estimation algorithm for bounded covariance distributions (Fact 12) to obtain an estimate $v \in \mathbb{R}^d$ with $\left\|v - \mathbf{E}_{X\sim P_{\theta^*}}[X]\right\|_2 \le O(\sqrt{\epsilon})$.

Let $\Omega = \{((\theta_{ij})_{1\le i<j\le d} \in \mathbb{R}^{d\times(d-1)/2}, (\theta_i)_{i\in[d]} \in \mathbb{R}^d \mid \max_{i\in[d]} \sum_{j=1}^{i-1} |\theta_{ji}| + \sum_{j=i+1}^{d} |\theta_{ij}| \le M, \max_{i\in[d]} |\theta_i| \le \alpha\}$. For any $\theta \in \Omega$, define $J(\theta)_{ij} = \theta_{ij}, \forall 1 \le i < j \le d$ and $h(\theta)_i = \theta_i + \sum_{j=1}^{i-1} \theta_{ji} v_j + \sum_{j=i+1}^{d} \theta_{ij} v_j, \forall i \in [d]$. Let $\Omega_{J,h} = \{(J(\theta), h(\theta)) \mid \theta \in \Omega\}$. Note that for any $\theta^1, \theta^2 \in \Omega$ and any $0 < \lambda < 1$, we have that $J(\lambda\theta^1 + (1-\lambda)\theta^2) = \lambda J(\theta^1) + (1-\lambda)J(\theta^2)$ and $h(\lambda\theta^1 + (1-\lambda)\theta^2) = \lambda h(\theta^1) + (1-\lambda)h(\theta^2)$, which implies that $\Omega_{J,h}$ is convex because of the convexity of Ω. Let $\theta \in \Omega$ and P_θ be the corresponding Ising distribution. We write P_θ in the "v-centered form", i.e., $P_\theta(x) = \frac{1}{Z(\theta)} \exp\left(\frac{1}{2}(x-v)^T J(\theta)(x-v) + h(\theta)^T x\right)$[3], where $Z(\theta)$ is the partition function. In this way, P_θ is an exponential family with sufficient statics $T(x) = ((x_i - v_i)(x_j - v_j)_{1\le i<j\le d}, (x_i)_{1\le i\le d})$.

Now we check the statements in Condition 6 one by one in order to apply Algorithm 1 to obtain an estimation of $J(\theta^*)$ and $h(\theta^*)$. By our choice of $\Omega_{J,h}$, we know that $\mathrm{diam}(\Omega_{J,h}) = O(d)$ and we can efficiently compute the projection of any point $z \in \mathbb{R}^{d\times(d-1)/2}$. From Fact 9, we can sample from P_θ within total variation distance γ in time $O(d(\log d + \log(1/\gamma)))$ for any $\gamma > 0$. Therefore the third statement holds. From Lemma 18, there is a universal constant $c > 0$ such that for any symmetric matrix $A \in \mathbb{R}^{d\times d}$ with zero diagonal and $b \in \mathbb{R}^d$, we have that

$$\mathbf{Pr}_{X\sim P_\theta}\left[|f(X) - \mathbf{E}[f(X)]| > t\right] \le 2\exp\left(-\frac{ct}{\sqrt{\|A\|_F^2 + \|b\|_2^2}}\right),$$

where $f(x) = (x-v)^T A(x-v) + b^T x, \forall x \in \{\pm 1\}^d$. This implies the second statement in Condtion 6. Moreover, From Theorem 19, we know that there is a universal constant $c' > 0$ such that for any symmetric matrix $A \in \mathbb{R}^{d\times d}$ with zero diagonal and $b \in \mathbb{R}^d$, we have that

$$\mathbf{Var}[(X-v)^T A(X-v) + b^T X] \ge c'(\|A\|_F^2 + \|b\|_2^2),$$

which implies the first statement in Condition 6. Then, we apply Algorithm 1 to obtain an estimation $\widehat{J}, \widehat{h}$ with $\sqrt{\|\widehat{J} - J(\theta^*)\|_F^2 + \|\widehat{h} - h(\theta^*)\|_2^2} \le O(\epsilon \log(1/\epsilon))$. Let $\widehat{\theta}_{ij} = \widehat{J}_{ij}, \forall i,j \in [d]$ and $\widehat{\theta}_i = \widehat{h}_i - \sum_{j=1}^{d} \widehat{J}_{ij} v_j$. From Theorem 7, we have that

$$d_{\mathrm{TV}}(P_{\widehat{\theta}}, P_{\theta^*}) \le O\left(\sqrt{\|\widehat{J} - J(\theta^*)\|_F^2 + \|\widehat{h} - h(\theta^*)\|_2^2}\right) \le O(\epsilon \log(1/\epsilon)),$$

where $P_{\widehat{\theta}}$ denotes the Ising model distribution corresponding to parameter $\widehat{\theta}$. In addition, by our algorithm, we have that $\max_{i\in[d]} \sum_{j\ne i} |\widehat{\theta}_{ij}| \le 1 - \eta$, and thus the output hypothesis satisfies Dobrushin's condition. ■

3. For simplicity, we also use $J(\theta)$ to note the $d \times d$ symmetric matrix with zero diagonal.