Learning Ising models from one or multiple samples

Learning Ising Models from One or Multiple Samples

Yuval Dagan
EECS & CSAIL, MIT
USA
dagan@mit.edu

Constantinos Daskalakis
EECS & CSAIL, MIT
USA
costis@csail.mit.edu

Nishanth Dikkala
GOOGLE RESEARCH & MIT
USA
nishanthd@google.com

Anthimos Vardis Kandiros
EECS & CSAIL, MIT
USA
kandiros@mit.edu

ABSTRACT

There have been two main lines of work on estimating Ising models: (1) estimating them from multiple independent samples under minimal assumptions about the model's interaction matrix ; and (2) estimating them from one sample in restrictive settings. We propose a unified framework that smoothly interpolates between these two settings, enabling significantly richer estimation guarantees from one, a few, or many samples.

Our main theorem provides guarantees for one-sample estimation, quantifying the estimation error in terms of the metric entropy of a family of interaction matrices. As corollaries of our main theorem, we derive bounds when the model's interaction matrix is a (sparse) linear combination of known matrices, or it belongs to a finite set, or to a high-dimensional manifold. In fact, our main result handles multiple independent samples by viewing them as one sample from a larger model, and can be used to derive estimation bounds that are qualitatively similar to those obtained in the afore-described multiple-sample literature. Our technical approach benefits from sparsifying a model's interaction network, conditioning on subsets of variables that make the dependencies in the resulting conditional distribution sufficiently weak. We use this sparsification technique to prove strong concentration and anti-concentration results for the Ising model, which we believe have applications beyond the scope of this paper.

CCS CONCEPTS

• **Theory of computation → Sample complexity and generalization bounds**; **Models of learning**; • **Mathematics of computing → Markov networks**.

KEYWORDS

Ising Model, Pseudo-Likelihood, Concentration Inequalities, Low Temperature, Single-Sample Estimation, Dependent Data

STOC '21, June 21–25, 2021, Virtual, Italy

ACM ISBN 978-1-4503-8053-9/21/06…$15.00
https://doi.org/10.1145/3406325.3451074

ACM Reference Format:
Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Anthimos Vardis Kandiros. 2021. Learning Ising Models from One or Multiple Samples. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21), June 21–25, 2021, Virtual, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3406325.3451074

1 INTRODUCTION

Markov Random Fields (MRFs) are a popular framework for representing high-dimensional distributions with conditional independence structure, represented via an undirected graph [39, 55]. The explicit representation of conditional independences allows for a more succinct representation of a distribution, decreasing the computational requirements to do inference. A special case of MRFs studied in this paper is the celebrated *Ising model* [34], which samples a binary vector, $x = (x_1, \ldots, x_n) \in \{\pm 1\}^n$, according to a measure of the following form:

$$\Pr_{J^*}[x] = \exp\left(x^\top J^* x/2 - F(J^*) - n\log 2\right), \tag{1}$$

where J^* is an $n \times n$ symmetric matrix with zero diagonal and $F(J^*) = \log\left(2^{-n}\sum_x \exp(x^\top J^* x/2)\right)$ is the so-called log-partition function. Notice that the term $J^*_{ij}x_i x_j$ in the exponent of the density encourages x_i and x_j to have equal or opposite signs depending on the sign and magnitude of J^*_{ij}, but this "local encouragement" can be overwritten by indirect interactions arising through paths between i and j in the undirected graph defined by the non-zero entries of J^*. Whenever i and j are disconnected in this graph, x_i and x_j are independent.

Since its introduction, the Ising model has found profound applications in a range of disciplines, including Statistical Physics, Computer Vision, Computational Biology, and the Social Sciences; see e.g. [13, 20, 22, 25, 26, 28]. These applications have motivated a long line of research aiming at estimating Ising models using samples. Some exciting progress on this front has appeared in recent years, including [11, 33, 35, 48, 50, 54, 57]. Importantly, most prior work assumes access to *multiple independent samples*, targeting estimating the interaction matrix J^* of a model under some conditions on J^*. Instead our focus in this work is to estimate Ising models from a *single* sample, which as we will shortly explain is a *more general problem*:

Single-Sample Ising Model Estimation: Given a family of interaction matrices $\mathcal{J} \subseteq \mathbb{R}^{n\times n}$ and one sample X from (1), where $J^* \in \mathcal{J}$, compute an estimate $\hat{J}(X)$ to minimize $\|\hat{J}(X) - J^*\|_F$.

Notice that estimating Ising models from one sample generalizes estimating them from multiple samples. This is because ℓ independent samples from an n-node Ising model with interaction matrix J^* can be viewed as one sample from an Ising model with $n\ell$ nodes, which belong to ℓ disconnected subnetworks that each have interaction matrix J^*.

Moreover, single-sample estimation is motivated by many applications where we may realistically only collect a single independent sample from a distribution. E.g., in applications of the Ising model in social network analysis, a sample from the model represents some binary behavior of the nodes in a social network, such as using an Android phone or an iPhone, voting for Democrats or Republicans, etc. In such applications, if we take a snapshot of the nodes' behaviors tomorrow, chances are that very little would change compared to their behavior today, and we certainly would not collect an independent sample. More broadly, lack of access to independent samples is ubiquitous in financial, meteorological, and geographical data, as well as social-network data [10, 43], where it has been studied in topics as diverse as criminal activity [31], welfare participation [6], school achievement [49], retirement plan participation [24], and obesity [17, 53]. Moreover, it has motivated a growing literature on single-sample statistical estimation, including [5, 7, 9, 12, 14, 15, 18, 21, 27, 30, 36–38, 41, 42, 44–47, 58].

Of course, one sample from (1) only carries n bits, while the matrix J^* to be estimated has $\Omega(n^2)$ real entries. Thus, one cannot hope to estimate J^* well from one sample without placing constraints on J^*. Said differently, the error in estimating J^* from one sample should depend on how complex J^* might be. This is the role played by $\mathcal{J}$ in the definition of our estimation problem. Our main result, presented shortly as Theorem 1, is that there exists an estimator whose error depends on the metric entropy of $\mathcal{J}$. Instantiating $\mathcal{J}$ in different ways, we obtain strong estimation guarantees when: (i) $\mathcal{J}$ is finite; (ii) it contains linear combinations of known matrices; (iii) it contains sparse linear combinations of known matrices; and (iv) it is a high-dimensional manifold. These are respectively Corollaries 1, 2, 3, and 4.

Prior to our work, the single-sample Ising model estimation literature had only studied quite restrictive special cases of our problem, namely the case $J^* = \beta J$, where J is a known matrix, and β is an unknown scalar strength parameter [9, 14], or slightly more general cases studied by follow-up work [9, 21, 30]. Restricted to this special case, our bounds provide quantitative improvements in the estimation error, as discussed in Section 2.4. However, our general theorem, as well as its corollaries in Settings (i)–(iv) discussed in the previous paragraph, provide vast extensions. E.g. (ii) and (iii) capture settings wherein we might know various social networks that individuals belong to (Facebook, LinkedIn, etc.) and expect that these all contribute to their behavior at different strengths. Setting (iv) captures settings of interest to Spatial Econometrics [2–4, 40] wherein we might be able to postulate a functional form for the interaction matrix and might be interested in estimating its parameters.

On the other hand, multiple-sample Ising model estimation is a widely studied problem with a long literature, going back to at least [16]. Yet, an efficient algorithm that learns Ising models on general (bounded-degree) graphs was only recently given in breakthrough work by [11], which has incited a renaissance of work on this topic [33, 35, 54, 57]. Since single-sample estimation generalizes multiple-sample estimation, as we have already discussed, our results for single-sample estimation allow us to obtain reconstruction guarantees for the following problem for any value of ℓ:

ℓ-Sample Ising Model Estimation: Given a family of interaction matrices $\mathcal{J} \subseteq \mathbb{R}^{n\times n}$ and ℓ independent samples from (1), where $J^* \in \mathcal{J}$, compute an estimate $\hat{J}$ to minimize $\|\hat{J} - J^*\|_F$.

Corollary 5 of Theorem 1 quantifies that access to multiple samples typically decreases the reconstruction error by a factor of $\widetilde{\Omega}(\sqrt{\ell})$. As such, we get reconstruction guarantees which smoothly interpolate between the single-sample estimation setting considered by [9, 14, 21, 30] and the more common $\omega(1)$-sample estimation setting considered by [11, 33, 35, 54, 57]. Interestingly, instantiating our result to the latter setting we obtain guarantees which are competitive to that work, as shown in Corollary 6 and the middle row of Table 1. Our sample complexity is typically higher, yet we derive it as a corollary of our main theorem which does not utilize independence between the samples. This further enables us to obtain similar bounds given two or more *dependent* samples as demonstrated by Corollary 7. See Table 1 for a summary of our results together with a comparison to prior work on estimation from a single and multiple samples.

Roadmap. In Section 2 we provide the statements of all of our results and briefly discuss their implications. In Section 3 we provide a sketch of the proof of our main result, namely Theorem 1, highlighting the challenges found in its proof and the innovations needed to overcome them. Finally, in Section 4 we review the most relevant results from the literature. We defer the complete details of the proof to the arXiv version of our paper[19].

2 OUR RESULTS

2.1 A General Upper Bound

In this section, we present a general upper bound that is a function of the covering numbers of the set $\mathcal{J}$, which represents the smallest number of elements from $\mathcal{J}$ that can approximate all elements of set $\mathcal{J}$. We begin with a definition.

Definition 1. *Given a normed space $(\mathcal{X}, \|\cdot\|)$, a set $\mathcal{V} \subseteq \mathcal{X}$ and $\epsilon > 0$, we say that a set $\mathcal{N} \subseteq \mathcal{V}$ is an ϵ-cover of $\mathcal{V}$ if for any $v \in \mathcal{V}$ there exists $u \in \mathcal{N}$ such that $\|u - v\| \le \epsilon$. The ϵ-covering number of $\mathcal{V}$ with respect to the norm $\|\cdot\|$, denoted by $N(\mathcal{V}, \|\cdot\|, \epsilon)$, is the minimum cardinality of an ϵ-cover.*

Our main result is stated below. As is standard in prior work, we parametrize our error in terms of a bound M on the infinity norm of the interaction matrices, $\|J\|_\infty = \max_i \sum_j |J_{ij}|$, which is called "width" in [35] and relaxes placing a bound on the maximum

Table 1: We state the estimation error $||\hat{J} - J^*||_F$ obtained by our work and prior work in different settings, ignoring some logarithmic factors. We present bounds under the standard assumption that $\|J^*\|_\infty$ is bounded by some constant M. Under this assumption, since $\|J\|_F \le \sqrt{n}\|J\|_\infty \le M\sqrt{n}$, a rate smaller than $M\sqrt{n}$ is non-trivial ; see Definition 1/Theorem 1.
*** In the first row, $f(\mathcal{J}, \epsilon) = \log N(\mathcal{J}, \|\cdot\|_2, \epsilon)$ is the metric entropy of family $\mathcal{J}$ under $\|\cdot\|_2$.**
**** In the last row, we consider the setting $\mathcal{J} = \{\beta J : |\beta| = O(1)\}$, where J is fixed and the estimation error is with respect to the parameter β. Here, $F(J^*)$ is the log-partition function, defined earlier.**

Family $\mathcal{J}$ of matrices	Single sample	ℓ samples
Arbitrary family $\mathcal{J}$ *	$\sqrt{f(\mathcal{J}, 1/n)}$ (Theorem 1)	$\sqrt{\frac{f(\mathcal{J}, 1/n\ell)}{\ell}}$ (Corollary 5)
Finite $\mathcal{J}$	$\sqrt{\log \lvert\mathcal{J}\rvert}$ (Corollary 1)	$\sqrt{\log \lvert\mathcal{J}\rvert/\ell}$ (Cor 1 & 5)
Linear combination of k known matrices	$\sqrt{k}$ (Corollary 2)	$\sqrt{k/\ell}$ (Cor 2 & 5)
s-sparse linear combination of k known matrices	$\sqrt{s \log k}$ (Corollary 3)	$\sqrt{s \log k/\ell}$ (Cor 3 & 5)
All matrices (unconstrained)	impossible	$n\sqrt{\log(n\ell)/\ell}$ (Corollary 6) $\sqrt{n}(\log n/\ell)^{1/4}$ [35]
Scalar multiples of a known matrix **	$1/\sqrt{F(J^*)}$ (Corollary 10) $1/\sqrt{F(J^*)}$ (under additional assumptions) [14] [9]	$1/\sqrt{\ell F(J^*)}$ (follows from our one-sample result)

degree [11, 54]. As shown in prior work [50], our single exponential dependence on M is necessary.[1]

Theorem 1. *Let $M > 0$ and let $\mathcal{J} \subseteq \{J : \|J\|_\infty \le M\}$ denote a collection of interaction matrices. There is an algorithm which, given a single sample $x \sim \Pr_{J^*}$ where $J^* \in \mathcal{J}$, outputs $\hat{J}$ such that with probability $\ge 1 - \delta$:*

$$\|\hat{J} - J^*\|_F \le C(M)\sqrt{\log N(\mathcal{J}, \|\cdot\|_2, 1/n) + \log(1/\delta) + \log\log n},$$

where $C(M)$ is an (single) exponential function of M and $\|\cdot\|_2$ denotes the spectral norm on matrices. Moreover, $\hat{J}$ is the minimizer over $\mathcal{J}$ of a convex function on the space of matrices, $\mathbb{R}^{n\times n}$. It can be computed in polynomial time if $\mathcal{J}$ is convex and projection onto $\mathcal{J}$ is efficiently computable.

Theorem 1 guarantees that we can find a matrix $\hat{J}$ that is close to the true interaction matrix J^* in Frobenius norm. In this general formulation, the error depends on the covering numbers of the set $\mathcal{J}$. In many interesting scenarios, the ϵ-cover of $\mathcal{J}$ will have size of the order of $(1/\epsilon)^k$, where k is a notion of dimension that is specific to each case. By applying Theorem 1, we obtain an error of the order of $\sqrt{k \log n}$ for constant M. If k is significantly less than n, this is a non-trivial bound, since both matrices $\hat{J}, J^*$ can have a Frobenius norm as high as $\Omega(\sqrt{n})$. We present examples where this is the case in the next section.

[1] While [50] provide a lower bound for multiple-sample estimation, their lower bound applies to our case as well because as we have explained single-sample estimation is more general than multiple-sample.

Remark 1 (Tightness of the bound)**.** *It is reasonable to expect that Theorem 1 is not completely tight. Tight upper bounds based on covering numbers are usually proved via the technique of chaining. However, technical difficulties arise once one tries to apply it in our scenario. Still, in all examples presented in the next section, this technique could remove only logarithmic factors, as our near-tight lower bounds provided in Section 2.3 establish.*

2.2 Applications of the Upper Bound

To showcase the power of Theorem 1, we now apply it to some concrete families $\mathcal{J}$. The families we consider capture both single-sample and multiple-sample Ising model estimation problems, in Sections 2.2.1 and 2.2.2 respectively. In all cases, we parametrize our bounds in terms of a bound M on the infinity norm of the matrices in $\mathcal{J}$ and a function $C(M)$ of which appears in our estimation error, as in Theorem 1.

2.2.1 Estimation from a Single Sample. The simplest case is when $\mathcal{J}$ is finite. Then, $N(\mathcal{J}, \|\cdot\|_2, \epsilon) \le |\mathcal{J}|$ for all $\epsilon \ge 0$ and we have:

Corollary 1. *If $\mathcal{J}$ is finite and all its elements J satisfy $\|J\|_\infty \le M$, our estimator satisfies*

$$\|\hat{J} - J^*\|_F \le C(M)\sqrt{\log |\mathcal{J}| + \log(1/\delta) + \log\log n},$$

with probability $\ge 1 - \delta$. Moreover, $\hat{J}$ can be computed in time $\mathrm{poly}(|\mathcal{J}|, n)$ *(i.e. polynomial time in $|\mathcal{J}|$ and n).*

Next, we consider settings where J^* is a linear combination of k known matrices, with unknown coefficients.

Corollary 2. *Let $J_1, \ldots, J_k$ be fixed matrices and let $\mathcal{J} = \{J = \sum_{i=1}^k \beta_i J_i : \|J\|_\infty \le M, \vec{\beta} \in \mathbb{R}^k\}$. Then, our estimator $\hat{J}$ satisfies $\|\hat{J} - J^*\|_F \le C(M)\sqrt{k \log n + \log(1/\delta)}$, with probability $\ge 1 - \delta$, and $\hat{J}$ can be computed in time* $\mathrm{poly}(n, k)$.

This can be extended to when J^* is a s-sparse linear combination of k known matrices, which enables us to obtain a bound with only a logarithmic dependence on k. For any $\vec{\beta} \in \mathbb{R}^k$ denote by $\|\vec{\beta}\|_0$ the number of nonzero coordinates of $\vec{\beta}$. The result is given below.

Corollary 3. *Let $J_1, \dots, J_k$ be fixed matrices, $s > 0$, and let $\mathcal{J} = \{J = \sum_{i=1}^k \beta_i J_i : \|J\|_\infty \le M, \|\vec{\beta}\|_0 \le s\}$. Then, our estimator $\hat{J}$ satisfies*

$$\|\hat{J} - J^*\|_F \le C(M)\sqrt{s(\log n + \log k) + \log(1/\delta)},$$

with probability $\ge 1-\delta$, and $\hat{J}$ can be computed in time $\mathrm{poly}(n,s)\cdot\binom{k}{s}$.

While Corollary 2 considers linear combinations of k known matrices, one can also consider non-linear settings, where, in general, the matrices lie in a k-dimensional manifold. We consider manifolds that are images of Lipschitz functions from convex subsets of $\mathbb{R}^k$ to the set of matrices. For this class, the following bound can be derived :

Corollary 4. *Let $h(\vec{\beta})$ be a function from $[-1,1]^k$ to the set of $n \times n$ matrices, that satisfies $\|h(\vec{\beta}) - h(\vec{\beta'})\|_2 \le L\|\vec{\beta} - \vec{\beta'}\|_\infty$ for some $L > 0$. Define $\mathcal{J} = \{J = h(\vec{\beta}) : \vec{\beta} \in [-1,1]^k, \|J\|_\infty \le M\}$. Then our estimator $\hat{J}$ satisfies*

$$\|\hat{J} - J^*\|_F \le C(M)\sqrt{k(\log n + \log L) + \log(1/\delta)},$$

with probability $\ge 1-\delta$.

2.2.2 Estimation from Several Samples. When we are given access to several independent or dependent samples, we can utilize them to obtain stronger guarantees. This is done via a reduction to the single-sample setting. As a first example, assume that ℓ independent samples from an n-dimensional Ising model are obtained. Notice that these can be viewed as a single sample from an $n\ell$ dimensional model. Thus, an application of Theorem 1 results in a gain of approximately $\sqrt{\ell}$ in the rate.

Corollary 5. *Let $M > 0$ and let $\mathcal{J} \subseteq \{J : \|J\|_\infty \le M\}$ denote a collection of interaction matrices. Assume that ℓ independent samples are obtained from $\Pr_{J^*}$ where $J^* \in \mathcal{J}$. There is an estimator $\hat{J}$ such that, with probability $\ge 1-\delta$,*

$$\|\hat{J} - J^*\|_F \le C(M)\sqrt{\frac{\log N(\mathcal{J}, \|\cdot\|_2, 1/(n\ell)) + \log(1/\delta) + \log\log n}{\ell}},$$

where the same comments for $C(M)$ and the complexity of computing $\hat{J}$ made in Theorem 1 apply.

Notice that Corollary 5 is phrased in terms of a general set $\mathcal{J}$. In particular, it can be applied to learn Ising models from multiple samples in the same setting studied by [35], where they learn J^* while only assuming that $\|J^*\|_\infty \le M$. Utilizing the fact that the space of interaction matrices is an $O(n^2)$-dimensional vector space, one obtains (similarly to Corollary 2):

Corollary 6. *Let $\mathcal{J} = \{J : \|J\|_\infty \le M\}$ and assume that ℓ independent samples from $\Pr_{J^*}$ where $J^* \in \mathcal{J}$ are obtained. Then, there is a polynomial time algorithm that finds $\hat{J} \in \mathcal{J}$ such that, w.p. $\ge 1-\delta$,*

$$\|\hat{J} - J^*\|_F \le C(M)\left(\sqrt{\frac{n^2\log(n\ell) + \log(1/\delta)}{\ell}}\right).$$

This provides a new polynomial-time algorithm for this problem. Compared to bound, [35] achieved an error of $\sqrt{n}(\log n/\ell)^{1/4}$, as also stated in Table 1.

Interestingly, as we discuss next, our results can be extended to settings where the samples are not independent.

Beyond Independent Samples. In many applications the learning task involves either a few or many dependent samples. For the sake of presentation, we assume time-series dependencies although other dependencies of a more complex structure can be studied in a similar fashion. Given an interaction matrix J_0 that controls the dependencies within each sample and J_1 that controls dependencies between consecutive samples, we define the following joint distribution over samples $x^1, \dots, x^\ell \in \{-1,1\}^n$:

$$\Pr_{J_0,J_1,\ell}\left[x^1 \cdots x^\ell\right] \propto \prod_{t=1}^{\ell} \exp\left(-(x^t)^\top J_0 x^t/2\right) \prod_{t=1}^{\ell-1} \exp\left(-(x^t)^\top J_1 x^{t+1}/2\right).$$

The following statement bounds the learning error, that can be meaningful even for $\ell = 2$:

Corollary 7. *Let $\ell \ge 2$, let $\mathcal{J}_0$ and $\mathcal{J}_1$ be collections of interaction matrices of infinity norm bounded by M, and let $(x^1, \dots, x^\ell) \sim \Pr_{J_0^*,J_1^*,\ell}$ for some $J_0^* \in \mathcal{J}_0$ and $J_1^* \in \mathcal{J}_1$. Then, there exists an estimator $(\hat{J}_0, \hat{J}_1)$ such that, w.p. $\ge 1-\delta$, both $\|J_0^* - \hat{J}_0\|_F$ and $\|J_1^* - \hat{J}_1\|_F$ are bounded by*

$$\frac{C(M)}{\sqrt{\ell}}\sqrt{C_0 + C_1 + \log\log n + \log(1/\delta)},$$

where

$$C_0 := \log N\left(\mathcal{J}_0, \|\cdot\|_2, \frac{1}{n\ell}\right)$$
$$C_1 := \log N\left(\mathcal{J}_1, \|\cdot\|_2, \frac{1}{n\ell}\right).$$

2.3 Lower Bounds

We first present a general lower bound based on the metric entropy of $\mathcal{J}$ and then we show that our lower bound is strong enough to provide nearly tight results for the cases of linear subspaces and finite sets. The following can be shown.

Theorem 2. *Let $r > 0$ and suppose there exists some $R, \alpha > 0$ and a family $\mathcal{J}$ of interaction matrices such that: (1) for all $J \in \mathcal{J}$ the infinity norm of J is bounded by $1-\alpha$ and the diameter[2] of $\mathcal{J}$ is bounded by R; and (2) it holds that*

$$\frac{\log N(\mathcal{J}, \|\cdot\|_F, 2r)}{2} \ge C(\alpha)R^2 + \log 2,$$

where $C(\alpha)$ is a specific constant determined in the proof. Then, any estimator $\hat{J}(x)$ based on a single sample attains a minimax error of $\max_{J^* \in \mathcal{J}} \mathbb{E}_{x\sim P_{J^*}}[\|\hat{J}(x) - J^*\|_F] \ge r/2$.

Using Theorem 2, one can derive a nearly-tight lower bound on the estimation error for linear combinations of k known matrices $J_1, \dots, J_k$:

[2] A set $\mathcal{K}$ has diameter at most R if for any $A, B \in \mathcal{K}$ we have $\|A - B\|_F \le R$.

Corollary 8. *Let $k \in \mathbb{N}$, let $J_1, \ldots, J_k$ be interaction matrices with disjoint supports*[3] *such that $\|J_i\|_\infty \le 1$ and $\|J_i\|_F \ge k$ for all i. Define $\mathcal{J} = \{J = \sum_i \alpha_i J_i : \alpha_i \in \mathbb{R}, \|J\|_\infty \le 1\}$. Then, any one-sample estimator $\hat{J}(x)$ has a minimax error of $\sup_{J^* \in \mathcal{J}} \mathbb{E}_{x \sim \Pr_{J^*}} \left[\|\hat{J} - J^*\|_F\right] \ge c\sqrt{k}$.*

In the proof of Corollary 8, one constructs a lower bound for a family of size $\exp(O(k))$. Hence, we derive the following tight lower bound of $\Omega(\sqrt{\log |\mathcal{J}|})$ on estimation from finite families of distributions:

Corollary 9. *Let $m > 0$. There exists a family $\mathcal{J}$ of cardinality $|\mathcal{J}| = m$ that satisfies $\sup_{J \in \mathcal{J}} \|J\|_\infty \le 1/2$, such that the minimax error satisfies $\max_{J^* \in \mathcal{J}} \mathbb{E}_{x \sim P_{J^*}} [\|\hat{J}(x) - J^*\|_F] \ge c\sqrt{\log m}$ (where $c > 0$ is a universal constant).*

2.4 Improved Bounds for Estimating a Single Parameter

We further present an application of our results to the single-sample setting studied in prior work [9, 14], namely estimating a single parameter β. The proof of the following statement follows from the main lemmas in the proof of Theorem 1.

Corollary 10. *Let $M > 0$, let J_0 be a fixed matrix with $\|J_0\|_\infty \le 1$ and let β^* be some unknown parameter satisfying $|\beta^*| \le M$. Then, there exists an estimator $\hat{\beta}$ from a single sample $x \sim \Pr_{\beta^* J_0}$ such that w.p. $\ge 1 - \delta$,*

$$|\hat{\beta} - \beta^*| \le C(M) F(\beta^* J_0)^{-1/2} \left(\log\log n + \log(1/\delta)\right),$$

where $F(\cdot)$ is defined as in (1).

Notice that the bound is inversely proportional to the square root of the partition function $F(J^*)$, which captures the strength of dependencies between the nodes and this bound is generally stronger than the one obtained using the Frobenius norm. Corollary 10 improves over prior work that required further assumptions to hold and obtained no guarantees at the vicinity of some phase transitions (see Section 4 for a comparison).

3 OVERVIEW OF TECHNIQUES

We start by presenting the main techniques used in this paper in Section 3.1 and proceed with a proof sketch in Section 3.2.

3.1 Key Technical Insights and Vignettes

From Low-Temperature to High-Temperature (Dobrushin). While nodes of the Ising model can be complexly dependent, when the correlations are sufficiently weak, the model shares important similarities to product measures. A well-studied mathematical formulation of weak dependencies for general random vectors is Dobrushin's uniqueness condition. For Ising models, a sufficient condition implying Dobrushin's is $\|J^*\|_\infty = \alpha < 1$, where α is a constant; see e.g. [23, 52].[4] While Dobrushin's condition implies multiple desirable properties (see e.g. [13, 56]), we will specifically use the fact that functions of the Ising model concentrate well under this condition; see e.g. [1, 13, 20, 29, 32]. Unfortunately, the regimes we are considering in this paper may lie well outside Dobrushin's condition, and the tools available to handle Ising models that do not satisfy Dobrushin's condition are significantly weaker and restricted, and concentration does not hold in general.

In this work, we prove concentration inequalities for Ising models outside of Dobrushin's condition via reductions to the Dobrushin regime: we show that we can condition on a subset of the variables, such that in the conditional distribution, the unconditioned variables satisfy Dobrushin. A basic example where we can see such behavior is when J is the incidence matrix of a bipartite graph, namely, there exists a set $I \subseteq [n]$ such that $J_{ij} = 0$ whenever either $i, j \in I$ or $i, j \in [n] \setminus I$. If we condition on $x_{-I} := x_{[n]\setminus I}$, then $\{x_i : i \in I\}$ are conditionally independent and particularly, satisfy Dobrushin. The following lemma generalizes this intuition. For the purposes of this lemma, we work with Ising models with *external fields*. Given an interaction matrix J^* and a vector h of external fields, we define the distribution over $x \in \{\pm 1\}^n$ by $\Pr_{J^*,h}(x) \propto \exp(x^{\mathrm{T}} J^* x/2 + h^T x)$.

Informal Lemma 1 (Conditioning Trick)**.** *Let $p_{J^*,h}(x)$ be an Ising model with interaction matrix J^* satisfying $\|J^*\|_\infty = M$ and any external field vector h. Then there exist $\ell = O(\log n)$ sets $I_1, \ldots, I_\ell \subseteq [n]$ such that:*

1. *Each $i \in [n]$ appears in exactly $\ell' = \lceil \ell/(16M) \rceil$ different sets I_j.*
2. *For all $j \in [\ell]$, the conditional distribution of x_{I_j}, conditioning on any setting of x_{-I_j}, satisfies Dobrushin's condition.*

We apply this lemma repeatedly in our proof, as it allows us to tap into the flexibility of dealing with weakly dependent random variables. As a first application, given a vector $a \in \mathbb{R}^n$, we obtain a lower bound on the variance of $a^\top x$. It is well known that if x is an *i.i.d.* vector of binary random variables, each with variance v, then $\mathrm{Var}(a^\top x) = v\|a\|_2^2$. Furthermore, if x satisfies Dobrushin's condition, then the entries of x are nearly independent and we can also show that $\mathrm{Var}(a^\top x) \ge \Omega(\|a\|_2^2)$. We will use Informal Lemma 1 to show that a similar lower bound holds even beyond Dobrushin's condition.

Informal Lemma 2 (Anti-Concentration)**.** *Suppose that x is sampled from an Ising model whose interaction matrix satisfies $\|J^*\|_\infty = O(1)$ and whose external field vector satisfies $\|h\|_\infty = O(1)$. Then, for all $a \in \mathbb{R}^n$,*

$$\mathrm{Var}(a^\top x) \ge \Omega(\|a\|_2^2).$$

Proof sketch. To prove this lemma, consider the sets $I_1, \ldots, I_\ell$ from Informal Lemma 1. First, we claim that there exists $j \in [\ell]$ such that $\|a_{I_j}\|_2^2 \ge \Omega(\|a\|_2^2)$. Indeed, by linearity of expectation, if we draw $j \in [\ell]$ uniformly at random then,

$$\mathbb{E}_j[\|a_{I_j}\|_2^2] = \mathbb{E}\left[\sum_{i=1}^n \mathbf{1}(i \in I_j) a_i^2\right] = \sum_{i=1}^n \frac{\ell'}{\ell} a_i^2 = \frac{\ell'}{\ell}\|a\|_2^2 \ge \Omega(\|a\|_2^2).$$

Hence, there exists a set I_j that achieves this expectation, namely, $\|a_{I_j}\|_2^2 \ge \Omega(\|a\|_2^2)$. Now using that, conditioning on x_{-I_j}, x_{I_j} has a low Dobrushin coefficient, as implied by Informal Lemma 1, we

[3] The support of a matrix J is defined as the set of its non-zero elements.

[4] Dobrushin's condition is slightly more general and defined in terms of a bound on the total influence exercised to any one node by the other nodes. However, as is often done in the literature, we use the slightly stronger but easier to interpret bound on $\|J^*\|_\infty$.

can bound $\mathrm{Var}[a^\top x \mid x_{-I_j}] \geq \Omega(\|a_{I_j}\|_2^2)$ as discussed above, using weak dependence. Since conditioning reduces the variance on expectation, we conclude that

$$\mathrm{Var}(a^\top x) \geq \mathbb{E}_{x_{-I_j}}[\mathrm{Var}[a^\top x | x_{-I_j}]] \geq \Omega(\|a_{I_j}\|_2^2) \geq \Omega(\|a\|_2^2).$$

□

Measure Concentration for Non-Polynomials. There are multiple recent works studying the concentration of polynomial functions of the Ising model [1, 20, 29, 32]. Here, we would like to bound the tails of general functions, in terms of their polynomial Taylor approximations. By a simple modification to the proof of [1], we can derive the following:

Theorem 3. *Let $f : \{0,1\}^n \mapsto \mathbb{R}$ be an arbitrary function and X be sampled from an Ising model which satisfies Dobrushin's condition. Then*

$$\Pr[|f(X) - \mathbb{E}f(X)| > t] \leq \exp\left(-c \min\left(\frac{t^2}{U_f}, \frac{t}{\max_x \|Hf(x)\|_2}\right)\right),$$

where

$$U_f := \|\mathbb{E}_X Df(X)\|_2^2 + \max_x \|Hf(x)\|_F^2.$$

Here $D_i f(x) = (f(x_{i+}) - f(x_{i-}))/2$ is the discrete derivative*, where x_{i+} and x_{i-} are obtained from x by replacing the value of x_i with 1 and -1, respectively. The vector of discrete derivatives is denoted by Df and Hf is the $n \times n$ matrix of second discrete derivatives.*

Theorem 3 can be trivially extended to derive bounds based on higher order Taylor expansion, extending [1, Theorem 2.2] for multi-linear polynomials.

3.2 Proof Sketch of Our Upper Bound

Using the tools from Section 3.1, we present a sketch of the proof of our main results. We start by describing the algorithm that is going to be used. A standard approach is *maximum likelihood estimation* (MLE), which outputs the maximizer $\hat{J}$ of the probability of the given sample x, namely, $\hat{J} := \mathrm{argmax}_J \Pr_J[x]$. Unfortunately, for Ising models, the MLE requires computing the partition function which is computationally hard to approximate [51]. A recourse, suggested by [14], is to compute the *maximum pseudo-likelihood estimator* (MPLE) of [7, 8] instead. One typically minimizes the negative log pseudo-likelihood,

$$\varphi(x; J) := -\sum_{i=1}^n \log \Pr_J[x_i \mid x_{-i}], \tag{2}$$

where $\Pr_J[x_i \mid x_{-i}]$ is the probability of $\Pr_J$ to draw x_i conditioned on the remaining entries of x, denoted x_{-i}. If $\mathcal{J}$ is a convex set, then this is a convex function which can be optimized using appropriate first-order optimization techniques to find an optimum $\hat{J}$.

A bound on the error can then be proved by the following steps. First, we show that for every $J_0 \in \mathcal{J}$ that is far from J^* we have

$$\varphi(x; J_0) \geq \varphi(x; J^*) + \Omega(1) \tag{3}$$

with high probability. One can prove this using a Taylor approximation of φ, while utilizing the first directional derivatives of φ that we define as

$$\frac{\partial \varphi(x; J)}{\partial A} := \lim_{t \to 0} \frac{\varphi(x; J + At) - \varphi(x; J)}{t}$$

and the second directed derivatives that we similarly define. Evaluating the Taylor approximation of $t \mapsto J^* + t(J_0 - J^*)$ at $t = 1$, one obtains that

$$\begin{aligned} \varphi(x; J_0) = \varphi(x; J^*) + \|J_0 - J^*\|_F \frac{\partial \varphi(x; J^*)}{\partial A} \\ + \frac{1}{2}\|J_0 - J^*\|_F^2 \frac{\partial^2 \varphi(x; J_x)}{\partial^2 A} \\ \text{, where } A = \frac{J_0 - J^*}{\|J_0 - J^*\|_F} \end{aligned} \tag{4}$$

and J_x is a point in the segment connecting J_0 with J^*. Hence, to show a large gap between $\varphi(x; J_0), \varphi(x; J^*)$ we need a good upper bound on the absolute value of the first derivative and a good lower bound on the second derivative.

We now turn to the specific challenges encountered when trying to prove these bounds. The first derivative takes the form

$$\frac{\partial \varphi(x; J^*)}{\partial A} = \sum_{i=1}^n \varphi_i'(x; J^*),$$

, where

$$\varphi_i'(x; J^*) := -\frac{\partial}{\partial A} \log \Pr_J[x_i \mid x_{-i}]\big|_{J=J^*}.$$

We notice that $\mathbb{E}[\varphi_i'(x; J^*) \mid x_{-i}] = 0$, hence it suffices to show concentration of the derivative around its mean to obtain a good upper bound. However, tail bounds on the gradient from prior work do not lead us to the optimal bound on the derivative in our setting. Instead, we use Lemma 1 to select a number of subsets $I_1, \ldots, I_\ell$ of $[n]$, such that conditioned on x_{-I_j}, x_{I_j} satisfies Dobrushin's condition. The lemma also guarantees that each $i \in [n]$ belongs to ℓ' different subsets I_j where ℓ' is a constant fraction of ℓ, which means we can write

$$\begin{aligned} \left|\frac{\partial \varphi(x; J^*)}{\partial A}\right| = \left|\sum_{i=1}^n \varphi_i'(x; J^*)\right| = \left|\frac{1}{\ell'} \sum_{j=1}^{\ell} \sum_{i \in I_j} \varphi_i'(x; J^*)\right| \\ \leq \frac{\ell}{\ell'} \max_j \left|\sum_{i \in I_j} \varphi_i'(x; J^*)\right| \\ \leq O\left(\max_{j \in [\ell]} \left|\sum_{i \in I_j} \varphi_i'(x; J^*)\right|\right). \end{aligned} \tag{5}$$

Hence, it suffices to bound each one of the terms that appear in the maximum. In fact, since each term $\sum_{i \in I_j} \varphi_i'(x; J^*)$ has zero mean conditioned on x_{-I_j}, it suffices to show that it concentrates around its expectation conditioned on x_{-I_j}. Given that conditioning on x_{-I_j}, x_{I_j} satisfies Dobrushin's condition, we can use the concentration inequality from Informal Theorem 3, to derive that

$$\left|\sum_{i \in I_j} \varphi_i'(x; J^*)\right| \leq O\left(\left\|\mathbb{E}\left[Ax \,\middle|\, x_{-I_j}\right]\right\|_2 + \|A\|_F\right),$$

with high probability. Applying (5) and union bounding over $j \in [\ell]$, we deduce that with high probability,

$$\left|\frac{\partial\varphi(x;J^*)}{\partial A}\right| \le \widetilde{O}\left(\max_{j\in[\ell]}\left\|\mathbb{E}\left[Ax \,\middle|\, x_{-I_j}\right]\right\|_2 + \|A\|_F\right). \tag{6}$$

We now show a lower bound on $\partial^2\varphi(x;J_x)/\partial^2 A$, where J_x is in the segment connecting J^* and J_0. Some simple calculations show that for every J in this segment,

$$\frac{\partial^2\varphi(x;J)}{\partial^2 A} \ge \Omega\left(\|Ax\|_2^2\right). \tag{7}$$

We then proceed by showing that: (a) the expectation of $\|Ax\|_2^2$ is lower bounded appropriately; and (b) it concentrates around its expectation. Note that (a) reduces to showing an expectation bound for a sum of squares of linear functions. This can also be phrased as a variance bound for linear functions of the Ising model, which is exactly the type of result that Informal Lemma 2 provides. Using it, we manage to prove that the expectation of the second derivative conditioned on x_{-I_j} is at least

$$\mathbb{E}\left[\|Ax\|_2^2\middle| x_{-I_j}\right] \ge \Omega\left(\left\|\mathbb{E}\left[Ax \,\middle|\, x_{-I_j}\right]\right\|_2^2 + \|A\|_F^2\right). \tag{8}$$

By concentration of polynomials under Dobrushin's condition [1], we will show that $\|Ax\|_2^2$ is at least the right hand side of (8) with high probability, and taking a union bound over $j \in [\ell]$, we derive that w.h.p.,

$$\begin{aligned}\frac{\partial^2\varphi(x;J_x)}{\partial^2 A} &\ge \|Ax\|_2^2\\ &\ge \Omega\left(\max_{j\in[\ell]}\left\|\mathbb{E}\left[Ax \,\middle|\, x_{-I_j}\right]\right\|_2^2 + \|A\|_F^2\right).\end{aligned} \tag{9}$$

If $\|J^* - J_0\|_F = \widetilde{\Omega}(1)$, we derive by (4), (6) and (9) that that inequality (3) holds w.h.p. Moreover, the further J_0 is from J^*, the higher is the probability.

We now have to use (3) to derive the error bound. To do that, we would like to show that for all J that are far from J^* in Frobenius norm, $\varphi(x;J) > \varphi(x;J^*)$. Since $\varphi(x;\hat{J}) \le \varphi(x;J^*)$, this would imply that $\hat{J}$ is close to J^*. Proving that this statement holds with high probability for all far enough points requires more than a union bound, since there might be infinitely many points. Instead, we will construct a finite subset $\mathcal{U}$ of these points such that every point is ϵ close to one in $\mathcal{U}$ ($\mathcal{U}$ forms an ϵ-net). By a union bound over $\mathcal{U}$ we prove that with high probability (3) holds for all points in $\mathcal{U}$. Since φ is Lipschitz as a function of the matrix J, this suffices to argue that for all far enough points, their φ value is much larger than that of J^*. We note that union bounding (3) over $|\mathcal{U}|$ events corresponding to all possible $J \in \mathcal{U}$, requires each event to hold with sufficiently high probability, and this holds whenever $\|J^* - J\|_F \ge \Omega(\sqrt{\log|\mathcal{U}|})$.

4 COMPARISON TO PRIOR WORK

Comparison with multiple-sample bounds. An important line of previous work focuses on learning Ising models from multiple independent samples. The first work that gives a polynomial-time algorithm for this problem is[11] and improved results were obtained by [33, 35, 54] and others. [35] showed that under the common assumption $\|J^*\|_\infty \le O(1)$, it is possible, using ℓ samples, to learn each row of J^* up to an error of $O((\log(n)/\ell)^{1/4})$, which translates to a Forbenius norm error of $O(n^{1/2}(\log(n)/\ell)^{1/4})$. In comparison, Corollary 5 can derive better guarantees even with one or a few samples, assuming additional structural assumptions on J^*. Further, Corollary 6 that assumes the same setting as [35], retains polynomial-time learnability, while reducing to a single-sample algorithm that does not utilize independence. This enables to consider dependent samples with only a small overhead.

Comparison with single-sample bounds. Another interesting line of work involves learning the Ising model from a single sample of the distribution. The first to work on this problem was [14], who assumed a single-parameter family, $\mathcal{J} = \{\beta J_0 \colon |\beta| \le M\}$ where the goal is to learn β. In subsequent work, [9] derived an improved bound and [30] presented an algorithm that jointly learns β and an *external field* θ, assuming that $\Pr[x] \propto \exp(-\beta x^\top Jx/2 + \theta\sum_i x_i)$. Further, [21] studied linear regression with Ising model dependencies, which corresponds to learning β together with multiple external field parameters. In comparison, Theorem 1 is the first to learn Ising models using one-sample from a complex family of matrices.

We further discuss the improvements over the prior work on single-sample estimation that are apparent in Corollary 10 and are essential for obtaining the results of this paper: (1) Removal of additional assumptions that require the log partition function $F(J^*)$ to be *well behaved*, yielding no guarantees in scenarios such as at the vicinity of some phase transitions. (2) Obtaining high probability estimates on single-parameter families that enables generalizing to arbitrary families via a union bound. These two improvements necessitates a new proof approach as presented in Section 3.

5 CONCLUSIONS AND FUTURE WORK

We obtained non-asymptotic rates for estimating the interaction matrix of an Ising model from a single sample, which are tight or near-tight in many interesting settings. An important feature of our analysis is that it also covers low temperature regimes, where our understanding is generally quite limited. A challenging problem for future investigation is obtaining estimation algorithms when the temperature is not bounded below. In this regime, we expect many interesting phenomena to arise, such as phase transitions, which might make the analysis more challenging. Another natural question is whether we could obtain similar results to ours without using a conditioning argument to handle the low temperature regime. This would further our understanding of the estimation problem and lead to a unified analysis for both low and high temperatures. Finally, building on this line of work, one could consider the task of performing statistical estimation when some of the nodes in the graph are not observed. In this setting we immediately lose some important properties of the full sample regime, such as the convexity of the log-likelihood function. As such, a broader set of tools might be required to tackle this more challenging setting.

ACKNOWLEDGEMENTS

We would like to thank Dheeraj Nagaraj for interesting discussions and help in proving the lower bound, and Frederic Koehler for the helpful discussion on the log partition function.

This work was supported by NSF Awards IIS-1741137, CCF-1617730 and CCF-1901292, by a Simons Investigator Award, by

the DOE PhILMs project (No. DE-AC05-76RL01830), by the DARPA award HR00111990021 and by the Onassis Foundation - Scholarship ID: F ZP 016-1/2019-2020.

REFERENCES

[1] Radosław Adamczak, Michał Kotowski, Bartłomiej Polaczyk, Michał Strzelecki, et al. 2019. A note on concentration for polynomials in the Ising model. *Electronic Journal of Probability* 24 (2019).

[2] Luc Anselin. 2001. Spatial econometrics. *A companion to theoretical econometrics* 310330 (2001).

[3] Luc Anselin. 2013. *Spatial econometrics: methods and models*. Vol. 4. Springer Science & Business Media.

[4] Luc Anselin and Raymond Florax. 2012. *New directions in spatial econometrics*. Springer Science & Business Media.

[5] Patrizia Berti, Irene Crimaldi, Luca Pratelli, Pietro Rigo, et al. 2009. Rate of convergence of predictive distributions for dependent data. *Bernoulli* 15, 4 (2009), 1351–1367.

[6] Marianne Bertrand, Erzo FP Luttmer, and Sendhil Mullainathan. 2000. Network effects and welfare cultures. *The Quarterly Journal of Economics* 115, 3 (2000), 1019–1055.

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

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

[9] Bhaswar B Bhattacharya and Sumit Mukherjee. 2018. Inference in Ising models. *Bernoulli* 24, 1 (2018), 493–525.

[10] Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. 2009. Identification of peer effects through social networks. *Journal of econometrics* 150, 1 (2009), 41–55.

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

[12] Guy Bresler and Dheeraj Nagaraj. 2018. Optimal single sample tests for structured versus unstructured network data. *arXiv preprint arXiv:1802.06186* (2018).

[13] Sourav Chatterjee. 2005. Concentration inequalities with exchangeable pairs (Ph. D. thesis). *arXiv preprint math/0507526* (2005).

[14] Sourav Chatterjee. 2007. Estimation in spin glasses: A first step. *The Annals of Statistics* 35, 5 (2007), 1931–1946.

[15] Justin Y. Chen, Gregory Valiant, and Paul Valiant. 2019. How bad is worst-case data if you know where it comes from? *arXiv* abs/1911.03605 (2019).

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

[17] Nicholas A Christakis and James H Fowler. 2013. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine* 32, 4 (2013), 556–577.

[18] Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Siddhartha Jayanti. 2019. Learning from Weakly Dependent Data under Dobrushin's Condition. In *Conference on Learning Theory*. 914–928.

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

[20] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. 2017. Concentration of multilinear functions of the Ising model with applications to network data. In *Advances in Neural Information Processing Systems*. 12–23.

[21] Constantinos Daskalakis, Nishanth Dikkala, and Ioannis Panageas. 2019. Regression from dependent observations. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 881–889.

[22] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. 2011. Evolutionary Trees and the Ising Model on the Bethe Lattice: A Proof of Steel's Conjecture. *Probability Theory and Related Fields* 149, 1 (2011), 149–189.

[23] RL Dobrushin and SB Shlosman. 1987. Completely analytical interactions: constructive description. *Journal of Statistical Physics* 46, 5-6 (1987), 983–1014.

[24] Esther Duflo and Emmanuel Saez. 2003. The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly journal of economics* 118, 3 (2003), 815–842.

[25] Glenn Ellison. 1993. Learning, Local Interaction, and Coordination. *Econometrica* 61, 5 (1993), 1047–1071.

[26] Joseph Felsenstein. 2004. *Inferring Phylogenies*. Sinauer Associates Sunderland.

[27] David Gamarnik. 2003. Extension of the PAC framework to finite and countable Markov chains. *IEEE Transactions on Information Theory* 49, 1 (2003), 338–345.

[28] Stuart Geman and Christine Graffigne. 1986. Markov Random Field Image Models and their Applications to Computer Vision. In *Proceedings of the International Congress of Mathematicians*. American Mathematical Society, 1496–1517.

[29] Reza Gheissari, Eyal Lubetzky, and Yuval Peres. 2017. Concentration inequalities for polynomials of contracting Ising models. *arXiv preprint arXiv:1706.00121* (2017).

[30] Promit Ghosal and Sumit Mukherjee. 2018. Joint estimation of parameters in Ising model. *arXiv preprint arXiv:1801.06570* (2018).

[31] Edward L Glaeser, Bruce Sacerdote, and Jose A Scheinkman. 1996. Crime and social interactions. *The Quarterly Journal of Economics* 111, 2 (1996), 507–548.

[32] Friedrich Götze, Holger Sambale, and Arthur Sinulis. 2019. Higher order concentration for functions of weakly dependent random variables. *Electronic Journal of Probability* 24 (2019).

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

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

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

[36] Aryeh Kontorovich and Maxim Raginsky. 2017. Concentration of measure without independence: a unified approach via the martingale method. In *Convexity and Concentration*. Springer, 183–210.

[37] Leonid Aryeh Kontorovich, Kavita Ramanan, et al. 2008. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability* 36, 6 (2008), 2126–2158.

[38] Vitaly Kuznetsov and Mehryar Mohri. 2015. Learning theory and algorithms for forecasting non-stationary time series. In *Advances in neural information processing systems*. 541–549.

[39] Steffen L Lauritzen. 1996. *Graphical models*. Vol. 17. Clarendon Press.

[40] James P LeSage. 2008. An introduction to spatial econometrics. *Revue d'économie industrielle* 123 (2008), 19–44.

[41] Ben London, Bert Huang, and Lise Getoor. 2016. Stability and generalization in structured prediction. *The Journal of Machine Learning Research* 17, 1 (2016), 7808–7859.

[42] Ben London, Bert Huang, Ben Taskar, and Lise Getoor. 2013. Collective stability in structured prediction: Generalization from one example. In *International Conference on Machine Learning*. 828–836.

[43] Charles F Manski. 1993. Identification of endogenous social effects: The reflection problem. *The review of economic studies* 60, 3 (1993), 531–542.

[44] Daniel J. McDonald and Cosma Rohilla Shalizi. 2017. Rademacher complexity of stationary sequences. *arXiv preprint arXiv:1106.0730* (2017).

[45] Mehryar Mohri and Afshin Rostamizadeh. 2009. Rademacher complexity bounds for non-iid processes. In *Advances in Neural Information Processing Systems*. 1097–1104.

[46] Mehryar Mohri and Afshin Rostamizadeh. 2010. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research* 11, Feb (2010), 789–814.

[47] Vladimir Pestov. 2010. Predictive PAC learnability: A paradigm for learning from exchangeable input data. In *Granular Computing (GrC), 2010 IEEE International Conference on*. IEEE, 387–391.

[48] Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. 2010. High-dimensional Ising model selection using l1-regularized logistic regression. *The Annals of Statistics* 38, 3 (2010), 1287–1319.

[49] Bruce Sacerdote. 2001. Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly journal of economics* 116, 2 (2001), 681–704.

[50] Narayana P Santhanam and Martin J Wainwright. 2012. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory* 58, 7 (2012), 4117–4134.

[51] Allan Sly and Nike Sun. 2014. Counting in two-spin models on d-regular graphs. *The Annals of Probability* 42, 6 (2014), 2383–2416.

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

[53] Justin G Trogdon, James Nonnemaker, and Joanne Pais. 2008. Peer effects in adolescent overweight. *Journal of health economics* 27, 5 (2008), 1388–1399.

[54] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. 2016. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems*. 2595–2603.

[55] Martin J Wainwright, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1, 1–2 (2008), 1–305.

[56] Dror Weitz. 2005. Combinatorial criteria for uniqueness of Gibbs measures. *Random Structures & Algorithms* 27, 4 (2005), 445–475.

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

[58] Bin Yu. 1994. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability* (1994), 94–116.