Learning Ising models from one or multiple samples

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

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

Nishanth Dikkala
GOOGLE RESEARCH & MIT
nishanthd@google.com

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

Abstract

There have been two separate lines of work on estimating Ising models: (1) estimating them from multiple independent samples under minimal assumptions about the model's interaction matrix [Bre15; Vuf+16; KM17; HKM17; WSD19]; and (2) estimating them from one sample in restrictive settings [Cha07; BM18; GM18; DDP19]. 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.

1 Introduction

Markov Random Fields (MRFs) are a popular framework for representing high-dimensional distributions with conditional independence structure, represented via an undirected graph [Lau96; WJ+08]. 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* [Isi25], 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. [GG86; Ell93; Fel04; Cha05; DMR11; DDK17]. 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 [SW12; RWL10; Bre15; Vuf+16; KM17; HKM17; WSD19]. 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 [Man93; BDF09], where it has been studied in topics as diverse as criminal activity [GSS96], welfare participation [BLM00], school achievement [Sac01], retirement plan participation [DS03], and obesity [CF13; TNP08]. Moreover, it has motivated a growing literature on single-sample statistical estimation, including [Bes74; Yu94; Cha07; KR+08; Ber+09; MR09; Pes10; MR10; Lon+13; KM15; LHG16; MS17; KR17; BM18; GM18; BN18; DDP19; Dag+19; CVV19].

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 [Cha07; BM18], or slightly more general cases studied by follow-up work [BM18; GM18; DDP19]. 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 [Ans01; LeS08; AF12; Ans13] 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 [CL68]. Yet, an efficient algorithm that learns Ising models on general (bounded-degree) graphs was only recently given in breakthrough work by [Bre15], which has incited a renaissance of work on this topic [Vuf+16; KM17; HKM17; WSD19]. 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 [Cha07; BM18; GM18; DDP19] and the more common $\omega(1)$-sample estimation setting considered by [Bre15; Vuf+16; KM17; HKM17; WSD19]. 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.

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) $\sqrt{n}(\log n/\ell)^{1/4}$ [KM17]
Scalar multiples of a known matrix **	$1/\sqrt{F(J^*)}$ (Corollary 10) $1/\sqrt{F(J^*)}$ (under additional assumptions) [Cha07; BM18]	$1/\sqrt{\ell F(J^*)}$ (follows from our one-sample result)

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 $\|J\|_F \leq \sqrt{n}\|J\|_\infty \leq M\sqrt{n}$, 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.

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 $\|u - v\| \leq \epsilon$. 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 [KM17] and relaxes placing a bound on the maximum degree [Bre15; Vuf+16]. As shown in prior work [SW12], our single exponential dependence on M is necessary.[1]

Theorem 1 (Follows from Theorem 4)**.** *Let $M > 0$ and let $\mathcal{J} \subseteq \{J : \|J\|_\infty \leq M\}$ 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 $\geq 1 - \delta$:*

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

[1]While [SW12] 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.

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.

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. The detailed statements and proofs are provided in Section 7.

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 $J_1, \dots, J_k$ be fixed matrices and let $\mathcal{J} = \{J = \sum_{i=1}^k \beta_i J_i \colon \|J\|_\infty \le M, \vec{\beta} \in \mathbb{R}^k\}$. 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 $\mathcal{J} = \{J = \sum_{i=1}^k \beta_i J_i \colon \|J\|_\infty \le M, \|\vec{\beta}\|_0 \le s\}$. 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 (see Section 7.1.4 for a general argument):

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

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 (Special case of Corollary 11)**.** *Let $M > 0$ and let $\mathcal{J} \subseteq \{J : \|J\|_\infty \leq M\}$ 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 $\geq 1 - \delta$,*

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

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 [KM17], where they learn J^* while only assuming that $\|J^*\|_\infty \leq M$. 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 $\mathcal{J} = \{J : \|J\|_\infty \leq M\}$ 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. $\geq 1 - \delta$,*

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

This provides a new polynomial-time algorithm for this problem. Comparing to our error bound, [KM17] 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 $x^1, \ldots, x^\ell \in \{-1,1\}^n$:

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

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

Corollary 7 (Special case of Corollary 12)**.** *Let $\ell \geq 2$, let $\mathcal{J}_0$ and $\mathcal{J}_1$ be collections of interaction matrices of infinity norm bounded by M, and let $(x^1, \ldots, x^\ell) \sim \Pr_{J_0^*,J_1^*,\ell}$ 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. $\geq 1-\delta$, both $\|J_0^* - \hat{J}_0\|_F$ and $\|J_1^* - \hat{J}_1\|_F$ are bounded by*

$$\frac{C(M)}{\sqrt{\ell}} \sqrt{\log N\left(\mathcal{J}_0, \|\cdot\|_2, \frac{1}{n\ell}\right) + \log N\left(\mathcal{J}_1, \|\cdot\|_2, \frac{1}{n\ell}\right) + \log\log n + \log(1/\delta)}.$$

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 is shown in Section 11.

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*

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

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 $\max_{J^* \in \mathcal{J}} \mathbb{E}_{x \sim P_{J^*}}[\|\hat{J}(x) - J^*\|_F] \geq r/2$.

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

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

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 $\sup_{J \in \mathcal{J}} \|J\|_\infty \leq 1/2$, such that the minimax error satisfies $\max_{J^* \in \mathcal{J}} \mathbb{E}_{x \sim P_{J^*}}[\|\hat{J}(x) - J^*\|_F] \geq c\sqrt{\log m}$ (where $c > 0$ is a universal constant).*

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

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

2.4 Improved bounds for estimating a single parameter

We further present an application to the single-sample setting studied in prior work [Cha07; BM18] on estimating a single parameter β (this follows from the main lemmas in the proof of Theorem 1):

Corollary 10. *Let $M > 0$, let J_0 be a fixed matrix with $\|J_0\|_\infty \leq 1$ and let β^* be some unknown parameter satisfying $|\beta^*| \leq M$. Then, there exists an estimator $\hat{\beta}$ from a single sample $x \sim \Pr_{\beta^* J_0}$ such that w.p. $\geq 1-\delta$, $|\hat{\beta} - \beta^*| \leq C(M)F(\beta^* J_0)^{-1/2}(\log\log n + \log(1/\delta))$ 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, defined formally in Section B. For Ising models, a sufficient condition implying Dobrushin's is $\|J^*\|_\infty = \alpha < 1$, where α is a constant; see e.g. [DS87; SZ92].[4] While Dobrushin's condition implies multiple desirable properties (see e.g. [Cha05; Wei05]), we will specifically use the fact that functions of the Ising model concentrate well under this condition; see e.g. [Cha05; DDK17; GLP17; GSS19; Ada+19]. 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 $\Pr_{J^*,h}(x) \propto \exp(x^{\mathrm{T}} J^* x/2 + h^T x)$.

[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. See Section B for the general form of the condition. However, as is often done in the literature, we use the slightly stronger but easier to interpret bound on $\|J^*\|_\infty$.

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 $\mathrm{Var}(a^\top x) \geq \Omega(\|a\|_2^2)$. 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 $a \in \mathbb{R}^n$,*

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

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 $\|a_{I_j}\|_2^2 \geq \Omega(\|a\|_2^2)$. Indeed, by linearity of expectation, if we draw $j \in [\ell]$ uniformly at random then,

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

Hence, there exists a set I_j that achieves this expectation, namely, $\|a_{I_j}\|_2^2 \geq \Omega(\|a\|_2^2)$. Now using that, conditioning on x_{-I_j}, x_{I_j} has a low Dobrushin coefficient, as implied by Informal Lemma 1, we 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 [DDK17; GLP17; GSS19; Ada+19]. 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 [Ada+19], 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*

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

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 [Ada+19, 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, $\hat{J} := \operatorname{argmax}_J \Pr_J[x]$. Unfortunately, for Ising models, the MLE requires computing the partition function which is computationally hard to approximate [SS14]. A recourse, suggested by [Cha07], is to compute the *maximum pseudo-likelihood estimator* (MPLE) of [Bes74; Bes75] 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

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

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 derivative $\varphi'(x; J^*)$ takes the form

$$\frac{\partial \varphi(x; J^*)}{\partial A} = \sum_{i=1}^{n} \varphi_i'(x; J^*) \quad ; \quad \varphi_i'(x; J^*) := -\frac{\partial}{\partial A} \log \Pr_J[x_i \mid x_{-i}]\big|_{J=J^*}.$$

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 $I_1, \ldots, I_l$ 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

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

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

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

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

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

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 [Ada+19], 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.,

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

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 $\varphi(x;\hat{J}) \leq \varphi(x;J^*)$, 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 $\|J^* - J\|_F \geq \Omega(\sqrt{\log|\mathcal{U}|})$.

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 Bresler [Bre15] and improved results were obtained by [Vuf+16; HKM17; KM17] and others. [KM17] showed that under the common assumption $\|J^*\|_\infty \leq O(1)$, 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 [KM17], 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 Chatterjee [Cha07], who assumed a single-parameter family, $\mathcal{J} = \{\beta J_0 \colon |\beta| \leq M\}$ where the goal is to learn β. In subsequent work, [BM18] derived an improved bound and Ghosal and Mukherjee [GM18] presented an algorithm that jointly learns β and an *external field* θ, assuming that $\Pr[x] \propto \exp(-\beta x^\top J x/2 + \theta \sum_i x_i)$. Further, Daskalakis et al. [DDP19] 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 Further Notation and Definitions

This section establishes the notational conventions used and presents some background definitions used throughout the paper. We start with the notational conventions.

Standard notations and definitions.

- Sets of indices: denote by $[n] = \{1, \ldots, n\}$. Given $I \subseteq [n]$, let $-I := [n] \setminus I$, and given $i \in [n]$ let $-i = -\{i\} = [n] \setminus \{i\}$.
- Indexed vectors and matrices: Given a vector $a = (a_1, \ldots, a_n)$ and a subset $I \subseteq [n]$, let a_I denote the $|I|$ coordinate vector$\{a_i \colon i \in I\}$. Similarly, for a matrix A of dimension $n \times m$, $I \in [n]$ and $I' \in [m]$, let $A_{II'}$ denote the corresponding submatrix. Similarly, let $A_I := A_{I[n]}$ and $A_{\cdot I} = A_{[n]I}$, and let $A_i := A_{\{i\}}$.
- Standard mathematical sets: Let $\mathcal{S}^{k-1}$ denote the k-dimensional unit sphere, $\{x \in \mathbb{R}^k \colon \|x\|_2 = 1\}$. And given $m, n > 0$ integers, let $M_{n\times m}(\mathbb{R}) := M_{n\times m}$ denote the space of real matrices of dimension $n \times m$.
- Ising model distributions: Given a symmetric matrix J with zeros on the diagonal, let $\Pr_J$ denote the Ising model with interaction matrix J, defined as in (1). We say that a random variable x with interaction matrix J and external field h is (M, γ)-bounded, if $\|J\|_\infty \leq M$, and if $\min_{i\in[n]} \mathrm{Var}(x_i|x_{-i}) \geq \gamma$ with probability 1. In other words, if for all x and all i, $\Pr[x_i = 1|x_{-i}](1 - \Pr[x_i = 1|x_{-i}]) \geq \gamma$.
- Absolute constants: We let the notations $C, c', C_1, \ldots$ denote constants that depend only on M and γ, and are bounded whenever M is bounded from above and γ from below (unless the dependence on γ and M is stated explicitly).
- Conditional variance: given two random variables X and Y, define $\mathrm{Var}[X|Y] := \mathbb{E}_X[(X - \mathbb{E}[X|Y])^2|Y]$. Since the conditional expectation $\mathbb{E}[X|Y]$ is a random variable which is a function of Y, so is $\mathrm{Var}[X|Y]$.

Matrix norms. Given a real matrix A of dimension $m \times n$, let $\|A\|_F^2 = \sum_{ij} A_{ij}^2$ denote the Frobenius norm, let

$$\|A\|_2 = \max_{u\in\mathbb{R}^n\setminus\{0\}} \frac{\|Au\|_2}{\|u\|_2}$$

and let

$$\|A\|_\infty = \max_{u\in\mathbb{R}^n\setminus\{0\}} \frac{\|Au\|_\infty}{\|u\|_\infty} = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |A_{ij}|.$$

The following inequalities are known for any symmetric matrix A: $\|A\|_2 \leq \|A\|_F$ and $\|A\|_2 \leq \|A\|_\infty$.

Dobrushin's condition. Next we define a variant of Dobrushin's uniqueness condition (high-temperature condition) for Ising models that we use. The more general form of the condition is presented in Section B.

Definition 2 (Dobrushin's condition)**.** *Given an Ising model x with interaction matrix J, we say that it satisfies* Dobrushin's condition *if $\|J\|_\infty < 1$, where $\alpha := \|J\|_\infty$ is called the* Dobrushin's coefficient*.*

Optimization over a vector space $\mathcal{V}$. Notice that replacing the interaction matrices $J_1, \ldots, J_k$ with other matrices $J'_1, \ldots, J'_k$ that span the same linear subspace of $M_{n\times n}(\mathbb{R})$ does not change the studied problem. Hence, we will forget about $J_1, \ldots, J_k$ and replace them with their span $\mathcal{V}$.

6 Analyzing the Maximum Pseudo-Likelihood Estimator

This section contains the proof of our main result which is the following theorem.

Theorem 4. *Let $\mathcal{J}$ be some set of matrices of infinity norm bounded by a constant M, and let*

$$\begin{aligned} R &= \min\left\{ r \geq 0 \colon r \geq C(M)\sqrt{\log\log n + \log N(\mathcal{J}, \|\cdot\|_2, cr^2/n) + \log(1/\delta)} \right\} \\ &\leq C(M)\sqrt{\log\log n + \log N(\mathcal{J}, \|\cdot\|_2, c/n) + \log(1/\delta)}. \end{aligned}$$

Then, with probability $1-\delta$, it holds that any point $J \in \mathcal{J}$ that satisfies $\|J - \hat{J}\|_F \geq R$, also satisfies $\varphi(J) \geq \varphi(J^*) + cR^2$. In particular, there exists an algorithm that, given one sample $x \sim \Pr_{J^*}$ where $J^* \in \mathcal{J}$, outputs $\hat{J} = \hat{J}(x)$ such that

$$\|\hat{J} - J^*\|_F \leq R.$$

Theorem 4 guarantees that we can find a matrix $\hat{J}$ that is R close to the true interaction matrix J^* in Frobenius norm. A detailed discussion about applications of Theorem 4, including the case where $\mathcal{J}$ is a k-dimensional linear subspace of $\mathbb{R}^{n\times n}$, is given in Section 7.

Section 6.1 contains an overview of the proof, while the main lemmas are presented in the following sections.

6.1 Overview of the proof

The algorithm used to estimate J^* will be the maximum pseudo-likelihood estimator (MPLE), as mentioned in Section 5:

$$\arg\max_{J\in\mathcal{J}} PL(J;x) := \arg\max_{J\in\mathcal{J}} \prod_{i\in[n]} \Pr_J[x_i|x_{-i}]. \tag{10}$$

In fact, the above maximization problem is concave and we are able to find an approximate solution using first-order methods, if $\mathcal{J}$ is a convex set. We will address the question of how to do this efficiently in Section A. For convenience in calculations, the function we will actually optimize is the negative log pseudo-likelihood:

$$\varphi(J) := -\log PL(J;x) = \sum_{i=1}^{n} (\log\cosh(J_i x) - x_i J_i x + \log 2). \tag{11}$$

A standard approach for showing consistency of the MPLE is by showing high probability upper bounds on the first derivatives of φ combined with a high probability lower bound on the smallest eigenvalue of the Hessian. To formalize this intuition in our setting, we begin by reviewing the definition of differentiation with respect to a matrix.

Definition 3. *If $f : M_{n\times n}(\mathbb{R}) \mapsto \mathbb{R}$ is a twice continuously differentiable function and $A \in M_{n\times n}(\mathbb{R}) \setminus \{0\}$, we define for all $J \in M_{n\times n}(\mathbb{R})$*

$$\frac{\partial f(J)}{\partial A} = \lim_{t\to 0} \frac{f(J+tA) - f(J)}{t} = \left.\frac{df(J+tA)}{dt}\right|_{t=0}.$$

In the case of the MPLE, some simple calculations show that

$$\frac{\partial \varphi(J)}{\partial A} = \frac{1}{2}\sum_{i=1}^{n}(A_i x)(\tanh(J_i x) - x_i); \quad \frac{\partial^2 \varphi(J)}{\partial A^2} = \frac{1}{2}\sum_{i=1}^{n}(A_i x)^2 \operatorname{sech}^2(J_i x). \tag{12}$$

The general strategy for proving that $\hat{J}$ is a good estimator of J^* is the following. First, since the MPLE finds the minimum of φ on the set $\mathcal{J}$, we know that $\varphi(\hat{J}) \leq \varphi(J^*)$. Next, we will show that any point J that is far from J^* in Frobenius norm should have a much larger value of φ than J^*. If we can show that this holds for all such matrices J with high probability, then certainly $\hat{J}$ should lie close to J^* in Frobenius norm, because $\varphi(\hat{J}) \leq \varphi(J^*)$. Hence, the bulk of the argument is in proving this gap between $\varphi(J)$ and $\varphi(J^*)$ if J is far from J^*. We will do this in two steps. First, we prove that this gap exists with high probability for a single $J \in \mathcal{J}$, namely, that $\varphi(J) > \varphi(J^*)$ with high probability. This is described in more detail in Section 6.1.1. The second step is to show that this holds for all J that are far from J^* with high probability. This requires finding a suitable ϵ-net of these matrices and taking a union bound over all the elements in this net. A crucial property in this step is the Lipschitzness of φ, which allows us to control the value of φ for all points that are close to a point in the net. The argument is described in Section 6.1.2.

6.1.1 A single dimensional problem

In this Section, we focus on a single $J_1 \in \mathcal{J}$. We would like to show that if $\|J_1 - J^*\|_F$ is large, then $\varphi(J_1) - \varphi(J^*)$ will also be large. This is formalized in the following Lemma.

Lemma 1. *Let $M > 0$, let x be drawn from the distribution parametrized by J^* and let $J_1 \neq J^*$ be such $\|J_1\|_\infty, \|J^*\|_\infty \leq M$. Then, there are constants $c, c' > 0$ depending only on M such that with probability $1 - \log n \exp(-c\|J^* - J_1\|_F^2)$, it holds that*

$$\varphi(J_1) \geq \varphi(J^*) + c'\|J_1 - J^*\|_F^2.$$

In order to prove this lemma, we have to somehow be able to compare $\varphi(J^*)$ with $\varphi(J_1)$. A common way to make this comparison is to view φ as a function on the line connecting J_1, J^*. Specifically, we define for all $t \in [0, \|J_1 - J^*\|_F]$,

$$g(t) = \varphi\left(J(t)\right) \quad \text{where } J(t) = J^* + t\frac{J_1 - J^*}{\|J_1 - J^*\|_F}.$$

Let $A = \frac{J_1 - J^*}{\|J_1 - J^*\|_F}$ and notice that it has a unit Forbenius norm and

$$g'(t) = \frac{\partial \varphi(J(t))}{\partial A}; \quad g''(t) = \frac{\partial^2 \varphi(J(t))}{\partial^2 A}.$$

Thus, the problem becomes 1-dimensional and to compare these two values, we can just use the Taylor expansion of g around J^*. Specifically, we have that

$$\begin{aligned}\varphi(J_1) = g(\|J_1 - J^*\|_F) &= g(0) + \|J_1 - J^*\|_F g'(0) + \frac{\|J_1 - J^*\|_F^2}{2} g''(\xi) \\ &= \varphi(J^*) + \|J_1 - J^*\|_F \frac{\partial \varphi(J^*)}{\partial A} + \frac{\|J_1 - J^*\|_F^2}{2} \frac{\partial^2 \varphi(J(\xi))}{\partial^2 A}\end{aligned}$$

where $\xi \in [0, \|J_1 - J^*\|_F$. Based on the previous expression, to show that $\varphi(J_1) > \varphi(J^*)$, we need an upper bound on $|\partial \varphi(J^*)/\partial A|$ and also, a lower bound on $\partial^2 \varphi(J)/\partial^2 A$ for all J in the segment connecting J_1, J^*. Moreover, we would like these two bounds to be comparable to each other, so that they can be combined in the Taylor formula to prove the desired inequality.

It would be desirable to prove such bounds with high probability. From the concentration inequalities literature on Ising models, we know that such inequalities hold when the model is in high temperature, namely, $\|J^*\|_\infty < 1$ (we need to extend some of these inequalities to apply to our case). However, our assumption is just that $\|J^*\|_\infty$ is bounded by a constant, which might be greater than 1. Hence, we need to reduce concentration in this more general case into concentration in high temperature. To do that, we first present a simple but powerful lemma that converts Ising models $\Pr_J$ with $1 < \|J\|_\infty \leq C$ to models satisfying Dobrushins condition, namely $\|J\|_\infty < 1$, by conditioning on a set of nodes. While in the first setting standard concentration inequalities are not guaranteed to hold and fundamental quantities of the distribution such as the partition function are generally hard to approximate, the second setting resembles the i.i.d. scenario. While this reduction is simple and has many limitations, it suffices to show that the learning rate obtained in the constant influence regime is at least as good as the optimal rate achievable under Dobrushin's condition. This optimality is made precise in Section 11.

Lemma 2. *Let $x = (x_1, \ldots, x_n)$ be an (M, γ)-Ising model, and fix $\eta \in (0, M]$. Then, there exist subsets $I_1, \ldots, I_\ell \subseteq [n]$ with $\ell \leq CM^2 \log n / \eta^2$ such that:*

1. *For all $i \in [n]$,*
$$|\{j \in [\ell] \colon i \in I_j\}| = \left\lceil \frac{\eta \ell}{8M} \right\rceil$$
.
2. *For all $j \in [\ell]$ and any value of x_{-I_j}, the conditional distribution of x_{I_j} conditioned on x_{-I_j} is an (η, γ)-Ising model.*

Furthermore, for any non-negative vector $\theta \in \mathbb{R}^n$ there exists $j \in [\ell]$ such that

$$\sum_{i \in I_j} \theta_i \geq \frac{\eta}{8M} \sum_{i=1}^n \theta_i.$$

The proof of Lemma 2 is presented in Section 8. It is a simple application of the probabilistic method and will be used multiple times throughout this work. We can now state the two lemmas bounding the first and the second derivatives of φ.

Lemma 3. *Let x be drawn from the distribution parametrized by J^* and $A \in \mathbb{R}^{n\times n}$. Assume $\|J^*\|_\infty \leq M$. Let $I_1, \ldots, I_\ell$ be the subsets obtained by Lemma 2 for $\eta = 1/2$. Then, there exist constants C, c depending only on M such that for any $t > 0$ we have*

$$\left|\frac{\partial\varphi(J^*)}{\partial A}\right| \leq t\left(\|A\|_F + \max_{j\in[l]} \|\mathbb{E}[Ax|x_{-I_j}]\|_2\right)$$

with probability at least

$$1 - C\log n \exp\left(-c\min\left(t^2, \frac{t\|A\|_F}{\|A\|_2}\right)\right).$$

Lemma 4. *Let x be drawn from the distribution parameterized by J^* and $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with zeros on the diagonal. Assume $\|J^*\|_\infty \leq M$. Let $I_1, \ldots, I_\ell$ be the subsets obtained by Lemma 2 for $\eta = 1/2$. Then, there exist constants C, c depending only on M such that for any $t > 0$ we have that*

$$\|Ax\|_2^2 \geq c\|A\|_F^2 + c\max_{j\in[l]} \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2^2 - t\left(\|A\|_F + \max_{j\in[l]} \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2\right)$$

with probability at least

$$1 - C\log n \exp\left(-c\frac{\min(t^2, t\|A\|_F)}{\|A\|_2^2}\right).$$

Consequently, for any $J \in \mathcal{J}$ we have

$$\Pr\left[\frac{\partial^2\varphi(J)}{\partial^2 A} < c'\|A\|_F^2 + c'\max_{j\in[l]} \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2^2\right] \leq C\log n \exp\left(-c\frac{\|A\|_F^2}{\|A\|_2^2}\right).$$

In both of those Lemmas, we use the technique of conditioning on subsets I_i defined in Lemma 2. Essentially, Lemma 4 states that with high probability, $\|Ax\|_2^2$ is lower bounded by some specific quantity. The same quantity appears as an upper bound for the first derivative in Lemma 3. This is not a coincidence. Indeed, we will easily show in the following sections that $\|Ax\|_2^2$ is a lower bound for $\partial^2\varphi(J)/\partial^2 A$ for all matrices J with bounded infinity norm. This means that Lemma 4 is essentially a lower bound of the second derivative. Thus, combining Lemmas 3 and 4, we can prove Lemma 1. This is done in Section 6.4.

Notice that the derivative is upper bounded by a value which is not constant, but it is rather bounded in terms of conditional expectations with respect to x_{-I_j}. Similarly, the second derivative is lower bounded in terms of the same quantity. Since in the Taylor expansion we are interested in the difference between the second and first derivatives, such a non-constant bound suffices to derive Lemma 1. Furthermore, one might not, in general, replace these bounds with constant bounds, since the term $\max_{j\in[\ell]} \|\mathbb{E}[Ax|x_{-I_j}]\|_2$ might not concentrate outside of Dobrushin's condition, and, $\partial\varphi(J^*)/\partial A$ and $\partial^2\varphi(J)/\partial A^2$ will fluctuate along with it. Next, we present the proofs of Lemmas 3 and 4.

6.1.2 Completing the proof for multiple J's

We will now explain how to use Lemma 1 to conclude the proof of Theorem 4. First of all, we know that the value of $\hat{J}$ is smaller than $\varphi(J^*)$, since the algorithm minimizes the negative log-pseudolikelihood φ. On the other hand, by Lemma 1, we know that if J is far from J^*, then with high probability it's value $\varphi(J)$ will be significantly larger than $\varphi(J^*)$. And the further J is, the higher is the probability. If we could prove that this holds with high probability for all points that are far from J^*, then we would be able to conclude that $\hat{J}$ is close to J^*. This suggests the following plan of attack: we should pick an $R > 0$ such that with high probability, all points J with $\|J - J^*\|_F \geq R$ satisfy $\varphi(J) > \varphi(J^*)$. This would imply that $\|J^* - \hat{J}\|_F \leq R$.

Suppose $\mathcal{A}_R$ is the set of points J with $\|J - J^*\|_F \geq R$. One obstacle in proving such a statement is that there is possibly an uncountable amount of matrices J that are R far from J^*. Hence, a simple union bound over these matrices would yield meaningful result. A common way to reason about uncountable families of objects is to define an ϵ-net on this set of objects. An ϵ-net of the set $\mathcal{A}_R$ is a finite subset of it, such that any point in $\mathcal{A}_R$ is ϵ-close to some point in the net. Clearly, a small ϵ means that the size of the ϵ-net will be larger. Suppose N is the size of the ϵ net. As a first step in our proof, we could try to prove that for all points J in the net, $\varphi(J) > \varphi(J^*) + \Omega(R)$. By Lemma 1 and a simple union bound, this happens with probability at least

$$1 - N \log n e^{-cR^2}$$

Notice that a larger R implies a higher probability that our claim is true. On the other hand, R will be our final bound for $\|J^* - \hat{J}\|_F$, so we would like to make it as small as possible. It is clear that for a sufficiently large probability of success we need $R = \Omega(\sqrt{\log N + \log\log n})$.

Suppose we chose such an R. Notice that we have only shown the desired claim for the points in the net. Our hope is that since the remaining points are close to the points in the net, their values will also be close. This property is characterized by the Lipschitzness of the function φ. Using some straightforward computations, we bound the Lipschitzness of φ by n in Section 6.4. This means that if $\epsilon = O(R/n)$, then all points in $\mathcal{A}_R$ will have a larger value of φ than J^*. Thus, the size $N = N(\mathcal{A}_R, \|\cdot\|_2, R/n)$ of the net will depend on R. Thus, in order to design a net that takes advantage of the Lipschitzness and is guaraneed to work with high probability, we should pick an R such that

$$R \geq \sqrt{\log N(R/n)}.$$

Since we would like to show that $\|J^* - \hat{J}\|_F$ is small, we would like to choose the smallest such R. Thus, our final error rate will be of the form

$$\inf\{R : R \geq \sqrt{\log N(\mathbb{E}, \|\cdot\|_2, R/n) + \log\log n}\}$$

This is exactly the guarantee that Theorem 4 gives us.

Section organization. Section 6.2 contains the proof of Lemma 3 that provides an upper bound for the derivative of the MPLE. In Section 6.3 we present the bound on the second derivative of the MPLE(Lemma 4). Lastly, in Section 6.4 we give a proof of Lemma 1 that compares the value of a single J to J^* and conclude with the proof of Theorem 4. Note that Lemma 2 will be presented in Section 8.

6.2 Bounding the Derivative of Log Pseudo-Likelihood

The central goal of this section is to prove Lemma 3. The general form of the derivative is a sum of tanh functions. The first thing one notices is that its mean is 0. Hence, it is enough to prove a strong enough concentration bound for this quantity. There are many challenges with this approach, which we will now present.

First, most concentration bounds that are known hold for Ising models that satisfy Dobrushin's condition. However, our model may or may not be in this state. Thus, we utilize Lemma 2 to find a small number of sets $I_j \subseteq [n]$, such that each $i \in [n]$ belongs to a small number of these sets. This allows us to write the derivative sum as a sum over all the sets I_j, thus focusing on the behavior of the function in each set. Lemma 2 also guarantees that if we condition on x_{-I_j}, the resulting Ising model will satisfy Dobrushin's condition. Hence, the problem reduces to bounding the terms of the sum belonging to I_j when we condition on x_{-I_j}.

This brings us to the next challenge, which involves the concentration bound. Most of the existing results on concentration inequalities for the Ising model focus on the case where the function is a multilinear polynomial. Since this is not the case for the derivative, we introduce a simple modification to the already existing techniques so that we obtain the desired concentration in our case. This will give us a bound that depends on the conditional expectation of a quadratic form conditioned on x_{-I_j}. This might seem insufficient at first, since we are not getting a uniform bound for the derivative, but rather one that depends on the values of x_{-I_j}. However, an analogous lower bound is proven for the strong convexity in Lemma 4. Hence, the two bounds match in all instantiations of x_{-I_j}, allowing us to complete the proof for the concentration of the derivative in a single direction A.

6.2.1 Proof of Lemma 3

First, we decompose the derivative according to terms corresponding to the sets $I_1, \ldots, I_\ell$ as specified by Lemma 2 for $\eta = 1/2$. Recall that each element $i \in [n]$ appears in exactly $\ell' = \lceil \eta \ell / 8 \rceil$ sets:

$$\left| \frac{\partial \varphi(J^*)}{\partial A} \right| = \frac{1}{2} \left| \sum_{i \in [n]} A_i x (x_i - \tanh(J_i^* x)) \right| \leq \frac{1}{\ell'} \sum_{j \in [\ell]} \left| \sum_{i \in I_j} A_i x (x_i - \tanh(J_i^* x)) \right| := \frac{1}{\ell'} \sum_{j \in [\ell]} |\psi_j(x; A)|. \tag{13}$$

We will bound the terms $\{\psi_j(x; A) \colon j \in [\ell]\}$ separately, and show that each term concentrates around zero. In order to do so, we will show that conditioned on any value for x_{-I_j}, $\psi_j(x; A)$ concentrates around zero, while the radius of concentration can depend on the specific value of x_{-I_j}. First, we would like to claim that this term is conditionally zero mean:

Claim 1. *For any $j \in [\ell]$, $\mathbb{E}\left[\psi_j(x; A) \;\middle|\; x_{-I_j} \right] = 0$.*

Proof. First, fix $i \in I_j$, and notice that since A and J^* have zeros on the diagonal, both $A_i x$ and $J_i x$ are constant conditioned on x_{-i}, hence

$$\mathbb{E}[A_i x (x_i - \tanh(J_i x)) | x_{-i}] = A_i x (\mathbb{E}[x_i | x_{-i}] - \tanh(J_i^* x)) = 0,$$

where the last equality follows from definition of the Ising mode. Next, notice that

$$\begin{aligned}\mathbb{E}\left[\psi_j(x;A) \;\middle|\; x_{-I_j}\right] &= \sum_{i\in I_j} \mathbb{E}\left[A_i x(x_i - \tanh(J_i^* x)) \;\middle|\; x_{-I_j}\right] \\ &= \sum_{i\in I_j} \mathbb{E}_{x_{I_j}}\left[\mathbb{E}\left[A_i x(x_i - \tanh(J_i^* x)) \mid x_{-i}\right]\right] = 0.\end{aligned}$$

□

Next, we will bound the radius of concentration of $\psi_j(x;A)$ conditioned on x_{-I_j}, and show that it is roughly bounded by $O\left(1 + \left\|\mathbb{E}\left[Ax \;\middle|\; x_{-I_j}\right]\right\|_2\right)$. Additionally, we derive concentration inequalities for the above term. In order to achieve this task, we utilize Lemma 2, which states that x_{I_j} is conditionally Dobrushin, conditioned on x_{-I_j}. Hence, we can utilize concentration inequalities for Dobrushin random variables, and achieve the following bound:

Lemma 5. *For any symmetric matrix A with zeros on the diagonal, we have*

$$\Pr\left[|\psi_j(x;A)| \geq t \;\middle|\; x_{-I_j}\right] \leq \exp\left(-c\min\left(\frac{t^2}{\|\mathbb{E}\left[Ax \;\middle|\; x_{-I_j}\right]\|_2^2 + \|A\|_F^2}, \frac{t}{\|A\|_2}\right)\right).$$

Most existing concentration bounds for the Ising model focus on multilinear polynomials, which is obviously not the form ψ_j has. Hence, in order to prove this bound, we modified slightly the existing proof of Theorem 5 to present a concentration inequality for general functions of ising models satisfying Dobrushin's condition, presented in Section 9. The concentration radius we get depend on bounds of the first and second discrete derivatives of the function. The full proof of Lemma 5 can be found in Section C.1.

Combining Lemma 5 with (13), we obtain the proof of Lemma 3, which we restate here for convenience.

Lemma 3. *Let x be drawn from the distribution parametrized by J^* and $A \in \mathbb{R}^{n\times n}$. Assume $\|J^*\|_\infty \leq M$. Let $I_1, \ldots, I_\ell$ be the subsets obtained by Lemma 2 for $\eta = 1/2$. Then, there exist constants C, c depending only on M such that for any $t > 0$ we have*

$$\left|\frac{\partial\varphi(J^*)}{\partial A}\right| \leq t\left(\|A\|_F + \max_{j\in[l]} \|\mathbb{E}[Ax|x_{-I_j}]\|_2\right)$$

with probability at least

$$1 - C\log n \exp\left(-c\min\left(t^2, \frac{t\|A\|_F}{\|A\|_2}\right)\right).$$

Proof. For any $j \in [\ell]$, Lemma 5 implies that

$$\Pr\left[|\psi_j(x;A)| \geq t\left(\|A\|_F + \left\|\mathbb{E}[Ax|x_{-I_j}]\right\|_2\right)\right] \leq \exp\left(-c\min\left(t^2, \frac{t\|A\|_F}{\|A\|_2}\right)\right).$$

By a union bound over $j \in [\ell]$, with probability at least $1 - C\log n\exp(-c\min(t^2, t\|A\|_F/\|A\|_2))$,

$$|\psi_j(x;A)| \leq t(\|A\|_F + \|\mathbb{E}[Ax|x_{-I_j}]\|_2)$$

holds for all $j \in [\ell]$. By (13), whenever this holds, we also have

$$\left|\frac{\partial \varphi(J^*)}{\partial A}\right| \leq \frac{1}{\ell'} \sum_{j \in [\ell]} |\psi_j(x; A)| \leq \frac{\ell}{\ell'} t \left(\|A\|_F + \max_{j \in [\ell]} \|\mathbb{E}[Ax | x_{-I_j}]\|_2 \right).$$

Since ℓ / ℓ' is at most a constant, the proof is complete. □

6.3 Strong Convexity of Log Pseudo-Likelihood

The central goal of this section is to prove Lemma 4. We break it into two parts. Section 6.3.1 deals with the overall proof of the lemma, while Section 6.3.2 contains an auxiliary lemma utilized in the proof. The approach here is quite similar to the one for the derivative. However, the specific tools differ, since this is an anticoncentration result. The argument begins by lower bounding the second derivative by a quadratic form depending only on the direction A of the derivative, namely $\|Ax\|_2^2$. Then, the problem reduces to showing anticoncentration for this quadratic form. This will be accomplished by establishing two claims.

First, we show that the mean of this form is bounded away from 0 by a quantity that depends on the Frobenius norm of Ax, conditional on the values x_{-I_j}. This would trivially be true if the spins were independent. In the case of the Ising model, we use Lemma 2 to find a subset of the nodes that has a small Dobrushin constant. This means that these nodes will be weakly correlated conditional on the rest, hence close to independent. This translates to the covariance matrix being diagonally dominant, from which the claim follows.

Second, we need to show that $\|Ax\|_2^2$ concentrates around its mean in a radius that is of the same or less order that the mean, conditional on x_{-I_j}. This concentration will be easier than the one obtained for the derivative, since we are now dealing with a quadratic function, to which known concentration results apply [Ada+19]. Therefore, we have shown that the second derivative in a particular direction A is sufficiently large.

6.3.1 Proof of Lemma 4

The first step will be to lower bound the second derivative by a quadratic form. This way, the task of proving strong convexity becomes significantly easier. For all $J \in \mathcal{J}$ and any direction A, we obtain that

$$\begin{aligned} 4\frac{\partial^2 \varphi(J)}{\partial A^2} &= \sum_{i=1}^{n} (A_i x)^2 \operatorname{sech}^2(J_i x) \geq \sum_{i=1}^{n} (A_i x)^2 \operatorname{sech}^2(\|Jx\|_\infty) \\ &= \|Ax\|_2^2 \operatorname{sech}^2(\|Jx\|_\infty) \geq \|Ax\|_2^2 \operatorname{sech}^2(\|J\|_\infty \|x\|_\infty) \qquad (14) \\ &\geq \|Ax\|_2^2 \operatorname{sech}^2(M), \qquad (15) \end{aligned}$$

where we used the facts that $\operatorname{sech} x \geq \operatorname{sech} y$ whenever $|x| \leq |y|$ and $\|J\|_\infty \leq M$ for all $J \in \mathcal{J}$. Notice that the bound we get only depends on the direction A of the derivative and not on the point J where we are calculating it. Hence, any bounds we manage to prove on this quantity will hold for all $J \in \mathcal{J}$. Our goal is to show that with high probability $\|Ax\|_2^2$ is sufficiently large.

The general strategy to showing this anticoncentration property will be to bound it's expectation away from 0 and then show that the function concentrates well around that value. We now focus on the first goal. Since $\|Ax\|_2^2$ is a sum of squares of linear functions, we proceed

with a lower bound on the second moment of any linear function. What we need essentially is a variance lower bound. Therefore, we have the following lemma.

Lemma 6. *Let y be an (M, γ)-Ising model. Then, for any vector $a \in \mathbb{R}^m$,*

$$\mathrm{Var}(a^\top x) \geq \frac{c\gamma^2 \|a\|_2^2}{M}.$$

The proof is presented in Subsection 6.3.2. Here now give a short outline of it.

Proof sketch. First, let's examine what would happen if all the coordinates of x were independent, each coordinate having a variance of at least γ. Then,

$$\mathrm{Var}(a^\top x) = \sum_i \mathrm{Var}(a_i x_i) = \sum_i a_i^2 \,\mathrm{Var}(x_i) \geq \gamma \|a\|_2^2.$$

We would like to apply the same logic in our situation. Specifically, we would like the coordinates of x to be as close to independent as possible. Therefore, we will use Lemma 2 to find a set of coordinates I such that x_I is an (α, γ)-Ising model conditioned on any value for x_{-I}, where $\alpha = c\gamma$. In this regime, the interactions between the nodes are weak, resulting in the covariance matrix of x_I conditioned on x_{-I} being diagonally dominant. This allows us to derive the same variance bound as if x_I was independent:

$$\mathrm{Var}\left[a^\top x \;\middle|\; x_{-I}\right] = \mathrm{Var}\left[a_I^\top x_I \;\middle|\; x_{-I}\right] \geq c\gamma \|a_I\|_2^2.$$

Lemma 2 guarantees that we can select the set I such that $\|a_I\|_2^2 \geq \alpha/(8M)\|a\|_2^2$, by substituting θ_i with a_i^2 in the lemma. Hence,

$$\mathrm{Var}\left[a^\top x \;\middle|\; x_{-I}\right] \geq c\gamma \|a_I\|_2^2 \geq \frac{c'\gamma\alpha\|a\|_2^2}{M} \geq \frac{c''\gamma^2\|a\|_2^2}{M}.$$

Finally, since conditioning can only decrease the conditional variance on expectation, the same bound holds for $\mathrm{Var}[a^\top x]$. □

Now, we would like to apply Lemma 6 to obtain a lower bound for the expectation of $\|Ax\|_2^2$. A simple application of the Lemma shows that:

$$\mathbb{E}[\|Ax\|_2^2] = \sum_i \mathbb{E}[(A_i x)^2] \geq \sum_i (\mathrm{Var}[(A_i x)] + \mathbb{E}[A_i x]^2) \tag{16}$$

$$\geq \sum_i \Omega(\|A_i\|_2^2) + \|\mathbb{E}[Ax]\|_2^2 = \Omega(\|A\|_F^2) + \|\mathbb{E}[Ax]\|_2^2. \tag{17}$$

Recall, however, that our goal in Lemma 4 is to lower bound the second derivative of φ in terms of the conditional expectation of Ax, conditioned on x_{-I_j}, for $j \in [\ell]$. The following lemma presents a variant of (16) when we condition on x_{-I_j}. Its proof is an application of Lemma 2 along with some simple calculations with conditional expectations.

Lemma 7. *Fix $A \in \mathcal{A}$. For any I_j and any x_{-I_j}, it holds that*

$$\mathbb{E}\left[\|Ax\|_2^2 \;\middle|\; x_{-I_j}\right] \geq c\gamma^2 \|A_{\cdot I_j}\|_F^2 + \left\|\mathbb{E}\left[Ax \;\middle|\; x_{-I_j}\right]\right\|_2^2.$$

Furthermore, there exists some $j \in [\ell]$ such that

$$\mathbb{E}\left[\|Ax\|_2^2 \;\middle|\; x_{-I_j}\right] \geq c\gamma^2/M$$

with probability 1.

Proof. Since conditioned on x_{-I_j}, x_{I_j} is a $(1/2, M)$ Ising model, we derive by Lemma 6 that

$$\begin{aligned}
\mathbb{E}\left[\|Ax\|_2^2 \,|x_{-I_j}\right] &= \sum_{i=1}^n \mathbb{E}\left[(A_i x)^2 | x_{-I_j}\right] = \sum_{i=1}^n \left(\mathbb{E}\left[A_i x | x_{-I_j}\right]^2 + \mathrm{Var}\left[A_i x | x_{-I_j}\right]\right) \\
&= \left\|\mathbb{E}\left[Ax|x_{-I_j}\right]\right\|_2^2 + \sum_{i=1}^n \mathrm{Var}\left[A_{i,I_j} x_{I_j} | x_{-I_j}\right] \geq \left\|\mathbb{E}\left[Ax|x_{-I_j}\right]\right\|_2^2 + \sum_{i=1}^n c\gamma^2 \left\|A_{i,I_j}\right\|_2^2 \\
&\geq \left\|\mathbb{E}\left[Ax|x_{-I_j}\right]\right\|_2^2 + c\gamma^2 \left\|A_{\cdot I_j}\right\|_F^2.
\end{aligned}$$

To prove the second part of the lemma, it suffices to show that there exists $j \in [\ell]$ such that $\|A_{\cdot I_j}\|_F^2 \geq c'\|A\|_F^2/M = c'/M$. Recall that the sets $\{I_j\}_{j\in[\ell]}$ were obtained from Lemma 2 with $\eta = 1/2$. By the last part of this lemma, if we substitute $\theta_i = \|A_{\cdot i}\|_2^2$, we derive that there exists a $j \in [\ell]$ such that

$$\|A_{\cdot I_j}\|_F^2 = \sum_{i \in I_j} \theta_i \geq \frac{c}{M} \sum_{i \in [n]} \theta_i = \frac{c\|A\|_F^2}{M} = \frac{c}{M},$$

recalling that $\|A\|_F = 1$ for all $A \in \mathcal{A}$. □

According to Lemma 7, if we can show that $\|Ax\|_2^2$ is concentrated at a radius of

$$O\left(\left\|A_{I_j}\right\|_F^2 + \left\|\mathbb{E}[Ax]|x_{-I_j}\right\|_2^2\right)$$

around its mean, conditional on x_{-I_j}, then anticoncentration follows. The next Lemma contributes to the proof in this direction.

Lemma 8. *Fix symmetric matrix A with zeros on the diagonal and $j \in [\ell]$. Then, for any fixed value of x_{-I_j} and any $t > 0$*

$$\Pr\left[\|Ax\|_2^2 < \mathbb{E}\left[\|Ax\|_2^2 \;\middle|\; x_{-I_j}\right] - t \;\middle|\; x_{-I_j}\right] \leq \exp\left(-\frac{c}{\|A\|_2^2} \min\left(\frac{t^2}{\|A\|_F^2 + \|\mathbb{E}\left[Ax \;\middle|\; x_{-I_j}\right]\|_2^2}, t\right)\right).$$

The proof is given in Section C.2. Note that by Lemma 2 we know that x_{I_j} is conditionally Dobrushin, conditioned on x_{-I_j}. The reason why Lemma 8 does not directly follow from the concentration inequality for quadratic forms (Theorem 5) is that the matrix $A^\top A$ does not necessarily have zeroes on the diagonal. Hence, the proof consists of a simple argument that reduces to the concentration of a polynomial of the form $x^\top \widetilde{A} x$ where $\widetilde{A}$ has zeros on the diagonal.

By combining Lemmas 7 and 8, we are now able to prove that $\|Ax\|_2^2$ is lower bounded with high probability, conditioned on x_{-I_j}. This allows us to complete the proof of Lemma 4.

Lemma 4. *Let x be drawn from the distribution parameterized by J^* and $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with zeros on the diagonal. Assume $\|J^*\|_\infty \le M$. Let $I_1, \ldots, I_\ell$ be the subsets obtained by Lemma 2 for $\eta = 1/2$. Then, there exist constants C, c depending only on M such that for any $t > 0$ we have that*

$$\|Ax\|_2^2 \ge c\|A\|_F^2 + c\max_{j\in[l]} \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2^2 - t\left(\|A\|_F + \max_{j\in[l]} \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2\right)$$

with probability at least

$$1 - C\log n \exp\left(-c\frac{\min(t^2, t\|A\|_F)}{\|A\|_2^2}\right).$$

Consequently, for any $J \in \mathcal{J}$ we have

$$\Pr\left[\frac{\partial^2\varphi(J)}{\partial^2 A} < c'\|A\|_F^2 + c'\max_{j\in[l]} \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2^2\right] \le C\log n \exp\left(-c\frac{\|A\|_F^2}{\|A\|_2^2}\right).$$

Proof. Let E_j denote the event that

$$\|Ax\|_2^2 > \mathbb{E}[\|Ax\|_2^2 \mid x_{-I_j}] - t\left(\|A\|_F + \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2\right). \tag{18}$$

By Lemma 8,

$$\Pr[E_j] = \mathbb{E}_{x_{-I_j}}\left[\Pr[E_j \mid x_{-I_j}]\right] \ge 1 - \exp\left(-c\min(t^2, t\|A\|_F)/\|A\|_2^2\right).$$

By a union bound, $\Pr\left[\bigcap_j E_j\right] \ge 1 - C\log n \exp\left(-c\min(t^2, t\|A\|_F)/\|A\|_2^2\right)$.

From now onward, assume that $\bigcap_j E_j$ holds. Lemma 7 implies that

$$\mathbb{E}\left[\|Ax\|_2^2 \mid x_{-I_j}\right] \ge c\|A_{\cdot I_j}\|_F^2 + \|\mathbb{E}[Ax|x_{-I_j}]\|_2^2$$

and additionally, that there exists some j such that $\|A_{\cdot I_j}\|_F^2 \ge c\|A\|_F^2$. We derive that

$$\|Ax\|_2^2 > c\|A_{\cdot I_j}\|_F^2 + \|\mathbb{E}[Ax|x_{-I_j}]\|_2^2 - t\left(\|A\|_F + \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2\right) \tag{19}$$

holds for all j. Let $j_1 = \max_j \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2$ and $j_2 = \max_j \|A_{\cdot I_j}\|_F$. Then, substituting j_1 in (19), we derive that

$$\|Ax\|_2^2 > \max_j \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2^2 - t\left(\|A\|_F + \max_j \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2\right).$$

By substituting j_2 in (19) we derive that

$$\|Ax\|_2^2 > c'\|A\|_F^2 - t\left(\|A\|_F + \max_j \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2\right).$$

Taking an average of the above two inequalities, we derive that

$$\|Ax\|_2^2 > c'\|A\|_F^2/2 + \max_j \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2^2/2 - t\left(\|A\|_F + \max_j \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2\right).$$

This concludes the proof for the first part of the lemma. For the second part, we substitute $t = c''\|A\|_F$ and use the inequality(14) for a lower bound on the second derivative. ∎

6.3.2 Proof of Lemma 6

We begin by proving Lemma 6. The idea is that if the total interaction strength of a node with all of its neighbors is small compared to the variance of each node, then the behavior is similar to that of independent samples. To prove that, we employ the following Lemma, the proof of which can be found in [Dag+19, Lemma 4.9], and is also a direct corollary of the earlier paper [BN+19, Theorem 3.1].

Lemma 9. *Let x be an (M, γ) Ising model with $M < 1$. Fix $i \in [n]$ and let $P_{x_{-i}|x_i=1}$ denote the conditioned distribution over x_{-i} conditioned on $x_i = 1$ and $P_{x_{-i}|x_i=-1}$ denote the conditioned distribution conditioned on $x_i = -1$. Then,*

$$W_1(P_{x_{-i}|x_i=1}, P_{x_{-i}|x_i=-1}) \leq \frac{2M}{1-M},$$

where W_1 is the ℓ_1-Wasserstein distance, namely $W_1(P, Q) = \min_\pi \mathbb{E}_{(x,y)\sim\pi}[\|x - y\|_1]$. The minimum is taken over all distributions π over $\{-1, 1\}^n \times \{-1, 1\}^n$, such that the marginals are P and Q, respectively.

This lemma essentially tells us that if M is small enough, then changing the spin of a single node is unlikely to influence the remaining ones. Using this fact, we can proceed as in the proof sketch of Section 6.3.1 to prove the claim. To prove it in the general case, we employ once more the conditioning trick to reduce the Dobrushin constant of the model. Specifically, Lemma 2 guarantees that we can find a subset I of nodes that are conditionally very weakly dependent, while still maintaining a constant fraction of the total variance of the linear function. The details are given below.

Proof of Lemma 6. Fix $a \in \mathbb{R}^d$. We first assume that $M \leq \gamma/4$, and then we prove it for all M. We start by arguing that for all $i \in [n]$,

$$\sum_{j\in[n]\setminus\{i\}} |\mathrm{Cov}(x_i, x_j)| \leq \frac{M}{1-M}.$$

The strategy will be to bound the Wasserstein distance between two Ising models conditional on different values of a single node. Using a tensorization argument, we have that if $X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_n)$ are random vectors over the same domain, then $W_1(X, Y) \geq \sum_{i\in[n]} W_1(X_i, Y_i)$. Applying this on the distributions $P_{x_{-i}|x_i=1}$ and $P_{x_{-i}|x_i=-1}$, we derive by Lemma 9 that

$$\begin{aligned}\sum_{j\in[n]\setminus\{i\}} \big|\mathbb{E}[x_j|x_i = 1] - \mathbb{E}[x_j|x_i = -1]\big| &= \sum_{j\in[n]\setminus\{i\}} W_1(P_{x_j|x_i=1}, P_{x_j|x_i=-1}) \\ &\leq W_1(P_{x_{-i}|x_i=1}, P_{x_{-i}|x_i=-1}) \leq \frac{2M}{1-M},\end{aligned}$$

where we also used the easily established property that for any two random variables U, V supported in a set of size 2, $W_1(U, V) = |\mathbb{E}[U] - \mathbb{E}[V]|$. We will use the above bound to get a

bound on $\mathrm{Cov}(x_i, x_j)$. Indeed,

$$\begin{aligned}
\mathrm{Cov}(x_i, x_j) &= \mathbb{E}[(x_i - \mathbb{E}x_i)x_j] = \mathbb{E}_{x_i}[(x_i - \mathbb{E}x_i)\mathbb{E}_{x_j}[x_j|x_i]] \\
&= \Pr[x_i = 1](1 - \mathbb{E}x_i)\mathbb{E}[x_j|x_i = 1] + \Pr[x_i = -1](-1 - \mathbb{E}x_i)\mathbb{E}[x_j|x_i = -1] \\
&= \frac{1}{2}(1 + \mathbb{E}[x_i])(1 - \mathbb{E}x_i)\mathbb{E}[x_j|x_i = 1] + \frac{1}{2}(1 - \mathbb{E}[x_i])(-1 - \mathbb{E}x_i)\mathbb{E}[x_j|x_i = -1] \\
&= \frac{1}{2}(1 - \mathbb{E}[x_i]^2)(\mathbb{E}[x_j|x_i = 1] - \mathbb{E}[x_j|x_i = -1]).
\end{aligned}$$

Combining the above inequalities, we obtain that

$$\sum_{j \in [n] \setminus \{i\}} |\mathrm{Cov}(x_i, x_j)| \leq \frac{\sum_{j \in [n] \setminus \{i\}} |\mathbb{E}[x_j|x_i = 1] - \mathbb{E}[x_j|x_i = -1]|}{2} \leq \frac{M}{1 - M}.$$

To conclude the proof for the setting where $M \leq \gamma/4$:

$$\begin{aligned}
\mathrm{Var}(a^\top x) = \mathbb{E}[a^\top (x - \mathbb{E}x)(x - \mathbb{E}x)^\top a] &= \sum_{i,j} a_{ij} \mathrm{Cov}(x_i, x_j) \geq \sum_{i=1}^n a_i^2 \mathrm{Var}(x_i) - \sum_{i \neq j} |a_i a_j \mathrm{Cov}(x_i, x_j)| \\
&\geq \sum_{i=1}^n a_i^2 \mathrm{Var}(x_i) - \sum_{i \neq j} \frac{(a_i^2 + a_j^2)|\mathrm{Cov}(x_i, x_j)|}{2} = \sum_{i=1}^n a_i^2 \left(\mathrm{Var}(x_i) - \sum_{j \neq i} |\mathrm{Cov}(x_i, x_j)| \right) \\
&\geq \sum_{i=1}^n a_i^2 \left(\gamma - \frac{M}{1 - M} \right) \geq \|a\|_2^2 \frac{\gamma}{2},
\end{aligned}$$

where we used the fact that $M \leq \gamma/4 < 1/2$, which implies that $M/(1 - M) \leq 2M \leq \gamma/2$.

In order to prove the Lemma without the above assumption, we again use the conditioning trick. Specifically, we find a subset of the nodes that satisfies Dobrushin's condition with a small enough M and then apply our result. To make this formal, notice that by Lemma 2, there exists a subset I of nodes such that conditioned on any x_{-I}, x_I is a $(\gamma/4, \gamma)$-Ising model. Moreover, it is guaranteed that

$$\sum_{i \in I} a_i^2 \geq \frac{c'\gamma}{M} \sum_{i \in [n]} a_i^2.$$

Hence, if we apply the previous result when conditioning on x_{-I} we get

$$\mathrm{Var}[a^\top x | x_{-I}] = \mathrm{Var}[a_I^\top x_I | x_{-I}] \geq \frac{\gamma \|a_I\|_2^2}{2} \geq \frac{c\gamma^2 \|a\|_2^2}{M}.$$

Since conditioning decreases the variance in expectation, we have that

$$\mathrm{Var}[a^\top x] \geq \mathbb{E}_{x_{-I}}[\mathrm{Var}[a^\top x | x_{-I}]] \geq \frac{c\gamma^2 \|a\|_2^2}{M},$$

as required. □

6.4 Proof of Lemma 1 and Theorem 4

The goal of this section is to prove the main Theorem 4. We do so in two steps. First, we use Lemmas 3 and 4 to complete the proof of Lemma 1. The strategy here is to use a lower bound on the Taylor approximation of φ around J_1 to prove that it's value will be larger than J^*. To prove the lower bound, we utilize the lower bound on $\frac{\partial^2 \varphi(J)}{\partial^2 A}$ fro all J and the upper bound of $\frac{\partial \varphi(J^*)}{\partial A}$. The precise argument is carried out in the following proof.

Lemma 1. *Let $M > 0$, let x be drawn from the distribution parametrized by J^* and let $J_1 \neq J^*$ be such $\|J_1\|_\infty, \|J^*\|_\infty \leq M$. Then, there are constants $c, c' > 0$ depending only on M such that with probability $1 - \log n \exp(-c\|J^* - J_1\|_F^2)$, it holds that*

$$\varphi(J_1) \geq \varphi(J^*) + c'\|J_1 - J^*\|_F^2.$$

Proof. Define the function $J\colon [0,1] \to \mathbb{R}$ by $J(t) = J^* + t(J_1 - J^*)$ such that $J(0) = J^*$ and $J(1) = J_1$. Let $A = J_1 - J^*$. Notice that

$$\frac{d\varphi(J(t))}{dt} = \frac{\partial \varphi(J)}{\partial A}\bigg|_{J=J(t)}; \quad \frac{d^2\varphi(J(t))}{dt^2} = \frac{\partial^2 \varphi(J)}{\partial A^2}\bigg|_{J=J(t)}.$$

Let $\mu = \min_{t\in(0,1)} \frac{d^2\varphi(J(t))}{dt^2}$ and let $r = \frac{d\varphi(J(t))}{dt}|_{t=0}$. Then, by the fundamental theorem of calculus,

$$\frac{d\varphi(J(t))}{dt} \geq r + t\mu.$$

This implies by the fundamental theorem of calculus that

$$\varphi(J(t)) \geq \varphi(J(0)) + tr + \frac{t^2\mu}{2}.$$

Substituting $t = 1$, we derive that

$$\varphi(J_1) \geq \varphi(J^*) + r + \frac{\mu}{2}.$$

From Lemma 4, with probability $1 - C\log n \exp\left(-c\frac{\|A\|_F^2}{\|A\|_2^2}\right)$, it holds that $\mu \geq c_0\|A\|_F^2 + c_0 \max_j \|\mathbb{E}[Ax \mid x_{-I_j}]\|_2^2$. By applying Lemma 3 with $t = O(\|A\|_F)$, we derive that with probability $1 - C\log n \exp\left(-c\frac{\|A\|_F^2}{\|A\|_2^2}\right)$, $|r| \leq \mu/4$. Whenever these two events hold, we have

$$\varphi(J_1) - \varphi(J^*) \geq r + \frac{\mu}{2} \geq \frac{-\mu}{4} + \frac{\mu}{2} \geq c_0\|A\|_F^2/4.$$

This holds with probability

$$1 - C\log n \exp\left(-c\frac{\|A\|_F^2}{\|A\|_2^2}\right) = 1 - C\log n \exp\left(-c\frac{\|J_1 - J^*\|_F^2}{\|J_1 - J^*\|_2^2}\right)$$

and notice that $\|J_1 - J^*\|_2 \leq \|J_1\|_2 + \|J^*\|_2 \leq \|J_1\|_\infty + \|J^*\|_\infty$ which is bounded by a constant. □

To complete the proof of Theorem 4, we need the bound of Lemma 1 to hold for all points $J \in \mathcal{J}$ uniformly. So far we have obtained a bound for comparing a single J with J^* that holds with probability proportional to $1 - \log ne^{-cr^2}$. Hence, by taking a union bound over a set of points of cardinality N, we obtain a probability concentration bound of the form $1 - N \log ne^{-cr^2}$. Since we would like this bound to hold for all points , we should choose this set so that it *covers* the set $\mathcal{J}$. A natural candidate for such a set is an *ϵ-net* of $\mathcal{J}$. Intuitively, it is a set of points such that every $J \in \mathcal{J}$ is close to at least one of them. Ideally, if two points are close under some notion of distance, we would like to conclude that the values of φ on these two points will also be close. This is exactly the content of the following auxiliary Lemma.

Lemma 10. *Let $M > 0$ and J_1, J_2 be two matrices. Then,*

$$|\varphi(J_1) - \varphi(J_2)| \leq n\|J_1 - J_2\|_2$$

Proof. We have that

$$\varphi(J) = \sum_{i=1}^{n} (\log \cosh(J_i x) - x_i J_i x + \log 2).$$

We can define the function $g : [0, 1] \mapsto \mathbb{R}$ by

$$g(t) = \varphi\left(tJ_1 + (1-t)J_2\right)$$

Clearly, g is computing the values of φ across the line segment connecting J_1, J_2. Hence, the desired inequality is equivalent to

$$|g(1) - g(0)| \leq n\|J_1 - J_2\|_2$$

Computing the derivative of g will thus give us a suitable bound for this difference. By the calculations done in the previous lemmas, it is clear that

$$g'(t) = \left.\frac{\partial \varphi(J)}{\partial A}\right|_{J = tJ_1 + (1-t)J_2}$$

where $A = J_1 - J_2$. We set $J(t) = tJ_1 + (1-t)J_2$. By the calculations done in previous parts of the paper, we have

$$\left.\frac{\partial \varphi(J)}{\partial A}\right|_{J = tJ_1 + (1-t)J_2} = \frac{1}{2} \sum_{i=1}^{n} \left((J_1 - J_2)_i x\right) \left(\tanh(J(t)_i x) - x_i\right)$$

Using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\left|\sum_{i=1}^{n} \left((J_1 - J_2)_i x\right) \left(\tanh(J(t)_i x) - x_i\right)\right| &\leq \sqrt{\sum_{i=1}^{n} \left((J_1 - J_2)_i x\right)^2 \sum_{i=1}^{n} \left(\tanh(J(t)_i x) - x_i\right)^2} \\ &= \|(J_1 - J_2)x\|_2 \sqrt{\sum_{i=1}^{n} \left(\tanh(J(t)_i x) - x_i\right)^2}\end{aligned}$$

Since the tanh funtion is bounded in $[-1, 1]$, we have

$$\sqrt{\sum_{i=1}^{n} \left(\tanh(J(t)_i x) - x_i\right)^2} \leq 2\sqrt{n}$$

Also, by the infinity norm bound of J_1, J_2 we have

$$\|(J_1 - J_2)x\|_2 \leq \|J_1 - J_2\|_2 \|x\|_2 = \sqrt{n}\|J_1 - J_2\|_2$$

We conclude that for any $t \in [0, 1]$

$$|g'(t)| \leq \frac{1}{2} 2\sqrt{n}\sqrt{n}\|J_1 - J_2\|_2 = n\|J_1 - J_2\|_2$$

By the mean value theorem, this implies

$$|g(1) - g(0)| \leq n\|J_1 - J_2\|_2$$

and the claim is complete.

□

We now use the preceding observations to conclude the proof of Theorem 4. The idea is to consider the set of points $\mathcal{A}_r$ whose distance from J^* is at least r in Frobenius norm. We would like to show that all of these points will have a value of φ that is r larger than J^* with high probability. To do this, we find an r/n-net of $\mathcal{A}_r$. By taking a union bound over this set, we ensure that the concentration bound will hold for all points in this set. Also, by the Lipschitzness argument, this implies that this holds for all points in $\mathcal{A}_r$. The failure probability is thus $N(r/n) \log n e^{-cr^2}$. We choose an r small enough to make this probability smaller than δ and this gives us the final guarantee.

Theorem 4. *Let $\mathcal{J}$ be some set of matrices of infinity norm bounded by a constant M, and let*

$$\begin{aligned} R &= \min\left\{ r \geq 0 \colon r \geq C(M)\sqrt{\log\log n + \log N(\mathcal{J}, \|\cdot\|_2, cr^2/n) + \log(1/\delta)} \right\} \\ &\leq C(M)\sqrt{\log\log n + \log N(\mathcal{J}, \|\cdot\|_2, c/n) + \log(1/\delta)}. \end{aligned}$$

Then, with probability $1 - \delta$, it holds that any point $J \in \mathcal{J}$ that satisfies $\|J - \hat{J}\|_F \geq R$, also satisfies $\varphi(J) \geq \varphi(J^*) + cR^2$. In particular, there exists an algorithm that, given one sample $x \sim \Pr_{J^*}$ where $J^* \in \mathcal{J}$, outputs $\hat{J} = \hat{J}(x)$ such that

$$\|\hat{J} - J^*\|_F \leq R.$$

Proof. Let $r > 0$. The idea is to find an ϵ-cover $\mathcal{N}$ for the set of elements $\mathcal{J}_{\geq r} := \{J \in \mathcal{J} \colon \|J^* - J\|_F \geq r\}$. Then, apply Lemma 1 for each element in this cover and argue by a union bound, that with high probability all elements in $\mathcal{N}$ exceed J^* in the cost function φ. Then, use Lipschitzness of this function φ to generalize from the net to $\mathcal{J}_{\geq r}$.

Let $\epsilon = c'r^2/(2n)$ where c' is the constant from Lemma 1, and we take $\mathcal{N}$ to be an ϵ-cover of $\mathcal{J}_{\geq r}$ with respect to the matrix norm $\|\cdot\|_2$.

By applying Lemma 1, each $J \in \mathcal{N}$ satisfies with probability $1 - C \log n \exp(-cr^2)$,

$$\varphi(J) \geq \varphi(J^*) + c'r^2.$$

By a union bound over the net, with probability at least $1 - |\mathcal{N}| \log n \exp(-cr^2)$, all these points satisfy the above inequality.

Next, assume that the above holds and we generalize from the cover to all of $\mathcal{J}_{\geq r}$. Take any point $J \in \mathcal{J}_{\geq r}$ and let J_0 be any point in $\mathcal{J}$ that satisfies $\|J_0 - J\|_2 \leq \epsilon$. Then, we have by Lemma 10 that

$$\varphi(J) \geq \varphi(J_0) - n\|J - J_0\|_2 \geq \varphi(J_0) - c'r^2/2 \geq \varphi(J^*) + c'r^2/2.$$

The failure probability is

$$|\mathcal{N}| \log n \exp(-cr^2) = N(\mathcal{J}_{\geq r}, \|\cdot\|_2, c'r^2/(2n)) \log n \exp(-cr^2) \leq N(\mathcal{J}, \|\cdot\|_2, c'r^2/(4n)) \log n \exp(-cr^2),$$

using the fact that for any sets $\mathcal{U} \subseteq \mathcal{U}$, any norm $\|\cdot\|$ and any $\epsilon > 0$, it holds that $N(\mathcal{U}', \|\cdot\|, \epsilon) \leq N(\mathcal{U}, \|\cdot\|, \epsilon/2)$.

Let $N = N(\mathcal{J}, \|\cdot\|_2, c'r^2/(4n))$ and notice that r has to be sufficiently large such that

$$N \log n e^{-cr^2} \leq \delta,$$

which holds whenever

$$\log N + \log\log n - cr^2 \leq \log \delta$$

or

$$r \geq C\sqrt{\log\log n + \log N + \log(1/\delta)}.$$

□

7 Applications

We begin with applications for learning from one sample, and then move to multiple samples.

7.1 Applications for learning from one sample

7.1.1 A collection of finite candidates

Assume that $\mathcal{J}$ is a finite set. Then, we have $N(\mathcal{J}, \|\cdot\|_2, 0) = |\mathcal{J}|$. In particular, we immediately obtain the bound of Corollary 1:

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

7.1.2 A Euclidean subspace of matrices

In this section, we assume that the matrix J^* is a linear combination of k known candidate matrices, $J_1, \ldots, J_k$, and further, that it has bounded infinity norm. In particular, define $\mathcal{J} = \{J : \sum_{i=1}^k \beta_i J_i, \|J\|_\infty \leq M, \beta_i \in \mathbb{R}\}$. Using standard arguments, we can obtain bounds on the covering numbers of this set:

Lemma 11. *Let $\mathcal{J} = \{J : \sum_{i=1}^k \beta_i J_i, \|J\|_\infty \leq M, \beta_i \in \mathbb{R}\}$, where $J_1, \ldots, J_k$ are known matrices. Then*

$$N(\mathcal{J}, \|\cdot\|_\infty, \epsilon) \leq \left(1 + \frac{2M}{\epsilon}\right)^k$$

Proof. Let $B(x, r)$ be the ball of center x and radius r w.r.t. $\|\cdot\|_\infty$ in the subspace spanned by $J_1, \ldots, J_k$. Then the set $\mathcal{J}$ is equal to $B(0, M)$. Let $|V|$ denote the volume of a subset V of this k-dimensional subspace. Suppose we cover $B(0, M)$ in the usual greedy way: each time we choose a ball $B(x, \epsilon/2)$ that is disjoint with the previous balls, where $x \in B(M)$ and we add it to the cover, until we cannot add any more balls. Let N be the number of centers selected using this process. Then clearly any point x in $B(0, M)$ is within ϵ distance from some center(otherwise the ball $B(x, \epsilon/2)$ could be added in the cover). Hence, these centers form an ϵ net of $B(0, M)$. Moreover, notice that all the selected balls are disjoint and fit inside the ball $B(0, M + \epsilon/2)$. Hence, their total volume is at most $|B(0, M + \epsilon/2)$. It follows that

$$N|B(0, \epsilon/2)| \leq |B(0, M + \epsilon/2)|$$

By scaling we know that

$$|B(0, \epsilon/2)| = \left(\frac{\epsilon}{2}\right)^k |B(0, 1)|$$
$$|B(0, M + \epsilon/2)| = \left(M + \frac{\epsilon}{2}\right)^k |B(0, 1)|$$

Notice that it doesn't matter which norm defines the balls. As long as the balls scale exponentially with the dimension, the volume does so too.

Consequently, we have that

$$N \leq \frac{|B(0, M + \epsilon/2)|}{|B(0, \epsilon/2)|} = \left(\frac{M + \epsilon/2}{\epsilon/2}\right)^k = \left(1 + \frac{2M}{\epsilon}\right)^k$$

□

By applying Theorem 4, we immediately obtain Corollary 2 :

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

Proof. For the statistical quarantee, apply Theorem 4, using the covering numbers from Lemma 11. The optimum can be found in polynomial time due to Lemma 21. □

7.1.3 Sparse estimation

Here, we assume that the true interaction matrix is a linear combination of just a few matrices in the subspace. Hence, we are dealing with ℓ_0 sparsity. Denote for any $\vec{\beta} \in \mathbb{R}^k$ by $\|\vec{\beta}\|_0$ the number of nonzero coordinates of $\vec{\beta}$. The goal is to have a learning rate that is dominated by the number of nonzero components of β. This is precisely the content of the following claim.

Corollary 3. *Let $J_1, \dots, J_k$ be fixed matrices, $s > 0$, and let $\mathcal{J} = \{J = \sum_{i=1}^k \beta_i J_i \colon \|J\|_\infty \le M, \|\vec{\beta}\|_0 \le s\}$. 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* $\mathrm{poly}(n, s) \cdot \binom{k}{s}$.

Proof. To prove the statistical guarantee, we need to bound the covering number of the set $\mathcal{J}$. To do that, we use the fact that $\mathcal{J}$ is a union of $\binom{n}{s}$ collections of matrices, each of them is an intersection of some s-dimensional subspace with the set of matrices of bounded infinity norm. By Lemma 11 the ϵ covering number of each such set $\mathcal{J}'$ is bounded by $N(\mathcal{J}', \|\cdot\|_\infty, \epsilon) \le \left(1 + \frac{2M}{\epsilon}\right)^s$. Taking the union over the ϵ coverings over all the $\binom{k}{s}$ collections, we get a covering for $\mathcal{J}$ of size at most $\left(1 + \frac{2M}{\epsilon}\right)^s \binom{k}{s}$. Hence, by plugging this number into the general rate of Theorem 4 we immediately obtain the result.

For the analysis of the runtime, notice that in order to optimize the MPLE, one has to iterate over all $\binom{k}{s}$ subspaces of dimension s and find the minimizer in each of those, each in polynomial time due to Corollary 2. □

7.1.4 High dimensional manifolds

There are settings where the interaction matrix lies in a nonlinear space. One such case is when the space is a k-dimensional manifold of matrices. In this case, we have some continuous mapping $h \colon D \to M_{n\times n}(\mathbb{R})$ where $D \subseteq \mathbb{R}^k$ is some domain and $\mathcal{J} = \{h(x) \colon x \in D, \|h(x)\|_\infty \le M\}$. We also assume that h is Lipschitz. With these assumption, we can obtain the following guarantee.

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 $\mathcal{J} = \{J = h(\vec{\beta}) \colon \vec{\beta} \in [-1, 1]^k, \|J\|_\infty \le M\}$. Then our estimator $\hat{J}$ satisfies $\|\hat{J} - J^*\|_F \le C(M)\sqrt{k(\log n + \log L) + \log(1/\delta)}$, with probability $\ge 1 - \delta$.*

Proof. Since h satisfies $\|h(x) - h(y)\|_2 \le L\|x - y\|$ for some $L > 0$, we have that for any set $D \subseteq \mathbb{R}^k$:

$$N(h(D), \|\cdot\|_2, \epsilon) \le N(D, \|\cdot\|, \epsilon/L).$$

In particular, if $D = [0, 1]^k$ and $\|\cdot\| = \|\cdot\|_\infty$ is the infinity norm over vectors in $\mathbb{R}^k$, then $N(D, \|\cdot\|_\infty, \epsilon) \le \epsilon^{-k}$. Hence, by a direct application of Theorem 4 we derive the following bound with probability $1 - \delta$:

$$\|\hat{J} - J^*\|_F \le C\sqrt{k \log(Ln) + \log(1/\delta)}.$$

□

7.2 Multiple independent samples

In this Section, we assume that we have access to multiple samples, either independent or correlated. An application of Theorem 4 gives us nontrivial bounds.

7.2.1 A general statement for independent samples

We start by giving a general bound for learning from ℓ independent samples. The following is a slightly more general formulation of Corollary 5.

Corollary 11. *Let $\mathcal{J}$ be a set of interaction matrices of infinity norm bounded by M. Assume that ℓ independent samples are obtained from $\Pr_{J^*}$ for some $J^* \in \mathcal{J}$. There is an estimator $\hat{J}$ for J^* such that for all $\delta > 0$, with probability $1 - \delta$,*

$$\begin{aligned}\|\hat{J} - J^*\|_F &\leq \frac{1}{\sqrt{\ell}} \min\left\{ r \geq 0 \colon r \geq C(M)\sqrt{\log N\left(\mathcal{J}, \|\cdot\|_2, \frac{cr^2}{n\ell}\right) + \log\log n + \log(1/\delta)} \right\} \\ &\leq \frac{C(M)}{\sqrt{\ell}} \sqrt{\log N\left(\mathcal{J}, \|\cdot\|_2, \frac{1}{n\ell}\right) + \log\log n + \log(1/\delta)}.\end{aligned}$$

Proof. In the setting of ℓ independent samples, we can view the concatenation of ℓ independent n-bit samples as a single $n\ell$-bit sample. Given an interaction matrix J of size $n \times n$, denote by $h(J)$ the interaction matrix of size $n\ell \times n\ell$ that corresponds to the joint distribution over ℓ samples. Formally, $h(J)$ is block-diagonal, that contains ℓ copies of J in the diagonal and zeros otherwise: $h(J)_{i+nj,i'+nj} = J_{i,i'}$ for any $i, i' \in [n]$ and $j \in \{0, \ldots, n-1\}$ and $h(J)_{i,j} = 0$ otherwise. We create the collection of interaction matrices that correspond to the joint distributions over ℓ-independent samples by $\mathcal{J}^\ell = \{h(J) \colon J \in \mathcal{J}\}$. In order to learn $J^* \in \mathcal{J}$ from ℓ samples, we will instead view it as learning a matrix $J^{\ell *} \in \mathcal{J}^\ell$ using a single sample. We will use this to get an estimate for J^*.

First of all, we bound the error of learning a matrix from J^ℓ. Start by noticing that $\|h(J)\|_\infty = \|J\|_\infty$, hence, assuming that $\|J\|_\infty \leq M$ for all $J \in \mathcal{J}$, we derive that $\|J^\ell\|_\infty \leq M$ for all $J^\ell \in \mathcal{J}^\ell$. Further, $\|h(J) - h(J')\|_2 = \|J - J'\|_2$ hence for any $\epsilon > 0$, we have $N(\mathcal{J}, \|\cdot\|_2, \epsilon) = N(\mathcal{J}^\ell, \|\cdot\|_2, \epsilon)$. Let $\hat{J}^\ell$ denote the estimator of $J^{\ell *} \in \mathcal{J}^\ell$ and the learning rate for learning rate is bounded by

$$\begin{aligned}\|\hat{J}^\ell - J^{\ell *}\|_F &\leq \min\left\{ r \geq 0 \colon r \geq C(M)\sqrt{\log N\left(\mathcal{J}^\ell, \|\cdot\|_2, \frac{cr^2}{n\ell}\right) + \log\log n + \log(1/\delta)} \right\} \\ &= \min\left\{ r \geq 0 \colon r \geq C(M)\sqrt{\log N\left(\mathcal{J}, \|\cdot\|_2, \frac{cr^2}{n\ell}\right) + \log\log n + \log(1/\delta)} \right\}.\end{aligned}$$

Lastly, notice $\|h(J) - h(J')\|_F = \sqrt{\ell}\|J - J'\|_F$, hence if we define $\hat{J} = h^{-1}(J^\ell)$ then we derive that

$$\|\hat{J} - J^*\|_F \leq \frac{1}{\sqrt{\ell}} \min\left\{ r \geq 0 \colon r \geq C(M)\sqrt{\log N\left(\mathcal{J}, \|\cdot\|_2, \frac{cr^2}{n\ell}\right) + \log\log n + \log(1/\delta)} \right\}.$$

This proves Corollary 11. ∎

7.2.2 Estimating the complete matrix from independent samples

Assume now that nothing is known about the matrix J, we have $\mathcal{J} = \{J \colon \|J\|_\infty \leq M\}$. Then, $\mathcal{J}$ is an intersection of a vector space of dimension bounded by n^2 with the set of matrices of bounded infinity norm. We can apply Corollary 11 in combination with the covering number bound of Lemma 11 to derive that

$$\|J^* - \hat{J}\|_F \leq C(M)\sqrt{n^2 \log(n\ell) + \log(1/\delta)},$$

where ℓ is the number of samples. Since the optimization is over a linear space, we derive a polynomial time algorithm. Hence, we proved the following.

Corollary 6. *Let $\mathcal{J} = \{J : \|J\|_\infty \leq M\}$ 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. $\geq 1-\delta$,*

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

7.2.3 Multiple samples with dependencies

We now extend the previous results in the case where the ℓ samples are not independent. Specifically, assume that one obtains ℓ samples, but instead of being independent, they have some dependency structure. We show how to formulate this as a one-sample learning problem. For the sake of presentation, we limit the discussion to time-series dependencies. However, similar arguments can handle general dependency structures. Assume that $x^1, \ldots, x^\ell$ are n-bit vectors, let J_0, J_1 denote two $n \times n$ matrices and define the joint distribution over $x^1, \ldots, x^\ell$ by

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

Here, J_0 is the interaction matrix that controls each sample x^t and J_1 controls the interaction between x^t and x^{t+1}. We would like to present this distribution as an Ising model on $n\ell$ random bits. The joint interaction matrix, that we denote by $h(J_0, J_1)$, is a block matrix with ℓ copies of J_0 on the diagonal and $\ell-1$ copies of J_1 on each of the sub-diagonals. We would like to estimate J_0^*, J_1^* from ℓ dependent samples, under the assumption that $J_0^* \in \mathcal{J}_0$ and $J_1^* \in \mathcal{J}_1$ for some collections of interaction matrices of bounded infinity norm. The details are as follows.

Corollary 12. *Let $M > 0$, let $\ell \geq 2$, let $\mathcal{J}_0$ and $\mathcal{J}_1$ be collections of interaction matrices of infinity norm bounded by M and let $(x^1, \ldots, x^\ell) \sim \Pr_{J_0^*,J_1^*,\ell}$ for some $J_0^* \in \mathcal{J}_0$ and $J_1^* \in \mathcal{J}_1$. Then, there exists an estimator $(\hat{J}_0, \hat{J}_1)$ that satisfies w.p. $1-\delta$,*

$$\max(\|J_0^* - \hat{J}_0\|_F, \|J_1^* - \hat{J}_1\|_F) \leq$$
$$\frac{1}{\sqrt{\ell}} \min\left\{r \geq 0 : r \geq C(M)\sqrt{\log N\left(\mathcal{J}_0, \|\cdot\|_2, \frac{cr^2}{n\ell}\right) + \log N\left(\mathcal{J}_1, \|\cdot\|_2, \frac{cr^2}{n\ell}\right) + \log\log n + \log(1/\delta)}\right\}$$
$$\leq \frac{C(M)}{\sqrt{\ell}}\sqrt{\log N\left(\mathcal{J}_0, \|\cdot\|_2, \frac{1}{n\ell}\right) + \log N\left(\mathcal{J}_1, \|\cdot\|_2, \frac{1}{n\ell}\right) + \log\log n + \log(1/\delta)}.$$

Proof. Let $\mathcal{J}_{01} = \{h(J_0, J_1) : J_0 \in \mathcal{J}_0, J_1 \in \mathcal{J}_1\}$ and we will reduce to the one-sample problem of estimating $J_{01}^* \in \mathcal{J}_{01}$. We start by computing some properties of $\mathcal{J}_{01}$. First of all, it has matrices of bounded infinity norm. Indeed, it is easy to see that $\|h(J_0, J_1)\|_\infty \leq \|J_0\|_\infty + 2\|J_2\|_\infty$. Secondly, we bound the covering numbers of $\mathcal{J}_{01}$. First, we bound the spectral norms of matrices $h(J_0, J_1)$ in terms of those of J_0 and J_1. Notice that for any $n\ell$ dimensional vector $\vec{u}$ denoted by $(u_i^t)_{t\in[\ell],i\in[n]}$

and we have

$$\vec{u}^\top h(J_0, J_1)\vec{u} = \sum_{t=1}^{\ell}(u^t)^\top J_0 u^t + \sum_{t=1}^{\ell-1}(u^t)^\top J_1 u^{t+1} \le \sum_{t=1}^{\ell}\|u^t\|_2^2\|J_0\|_2 + \sum_{t=1}^{\ell-1}\|(u^t)\|_2\|J_1\|_2\|u^{t+1}\|_2$$
$$\le \|J_0\|_2\sum_{t=1}^{\ell}\|u^t\|_2^2 + \|J_1\|_2\sum_{t=1}^{\ell-1}(\|(u^t)\|_2^2 + \|u^{t+1}\|_2^2)/2 \le (\|J_0\|_2 + \|J_1\|_2)\|\vec{u}\|_2^2.$$

where we used the arithmetic and geometric means inequality. This implies that $\|h(J_0, J_1)\|_2 \le \|J_0\|_2 + \|J_1\|_2$. Since h is a linear function, we have that for any J_0, J_1, J_0', J_1', $\|h(J_0, J_1) - h(J_0', J_1')\|_2 = \|h(J_0 - J_0', J_1 - J_1')\|_2 \le \|J_0 - J_0'\|_2 + \|J_1 - J_1'\|_2$. This implies that for any $\epsilon > 0$, one has

$$N(\mathcal{J}_{01}, \|\cdot\|_2, \epsilon) \le N(\mathcal{J}_0, \|\cdot\|_2, \epsilon/2)N(\mathcal{J}_1, \|\cdot\|_2, \epsilon/2).$$

Indeed, given an $\epsilon/2$-cover $\mathcal{N}_0$ for $\mathcal{J}_0$ and $\mathcal{N}_1$ for $\mathcal{J}_1$, we can take $\mathcal{N}_{01} = \{h(J_0, J_1)\colon J_0 \in \mathcal{N}_0, J_1 \in \mathcal{N}_1\}$, and it forms an ϵ-cover for $\mathcal{J}_{01}$. We derive that $J_{01}^* \in \mathcal{J}_{01}$ can be estimated with an error of

$$\|J_{01}^* - \hat{J}_{01}\|_F \le$$
$$\min\left\{r \ge 0\colon r \ge C(M)\sqrt{\log N\left(\mathcal{J}_0, \|\cdot\|_2, \frac{cr^2}{n\ell}\right) + \log N\left(\mathcal{J}_1, \|\cdot\|_2, \frac{cr^2}{n\ell}\right) + \log\log n + \log(1/\delta)}\right\}.$$

Next, we translate back to guarantees on estimating J_0^* and J_1^*. Notice that

$$\|h(J_0, J_1) - h(J_0', J_1')\|_F^2 = \ell\|J_0 - J_0'\|_F^2 + 2(\ell - 1)\|J_1 - J_1'\|_F^2.$$

This immediately implies the rate given in the statement. □

8 The Conditioning Trick

The main purpose of this section is to prove the following Lemma, which is used multiple times in the proof of Theorem 4.

Lemma 2. *Let $x = (x_1, \ldots, x_n)$ be an (M, γ)-Ising model, and fix $\eta \in (0, M]$. Then, there exist subsets $I_1, \ldots, I_\ell \subseteq [n]$ with $\ell \le CM^2\log n/\eta^2$ such that:*

1. *For all $i \in [n]$,*
$$|\{j \in [\ell]\colon i \in I_j\}| = \left\lceil \frac{\eta\ell}{8M} \right\rceil$$
.
2. *For all $j \in [\ell]$ and any value of x_{-I_j}, the conditional distribution of x_{I_j} conditioned on x_{-I_j} is an (η, γ)-Ising model.*

Furthermore, for any non-negative vector $\theta \in \mathbb{R}^n$ there exists $j \in [\ell]$ such that

$$\sum_{i \in I_j}\theta_i \ge \frac{\eta}{8M}\sum_{i=1}^{n}\theta_i.$$

Intuitively, this lemma says that we can select a small number of subsets, such that each element $i \in [n]$ is contained in at least a constant fraction of these sets. At the same time, we want the variables in these subsets to be weakly dependent when we condition on the rest. The proof relies on the following technical lemma.

Lemma 12. *Let A be a matrix with zero on the diagonal, and $\|A\|_\infty \leq 1$. Let $0 < \eta < 1$. Then, there exist subsets $I_1, \ldots, I_\ell \subseteq [n]$, with $\ell \leq C \log n / \eta^2$ such that:*

1. *For all $i \in [n]$,*
$$|\{j \in [\ell] \colon i \in I_j\}| = \lceil \eta \ell / 8 \rceil \tag{20}$$
2. *For all $j \in [\ell]$ and all $i \in I_j$,*
$$\sum_{k \in I_j} |A_{ik}| \leq \eta \tag{21}$$

This Lemma lies in the heart of the proof. It essentially ensures that we can select subsets of nodes that conditionally satisfy Dobrushin's condition with an arbitrarily small constant. It also shows that these subsets can cover all the nodes. Thus, it allows breaking up a sum over all the nodes to a sum over all these subsets with only a logarithmic error factor. The proof is an application of the probilistic method, using a simple, but strong technique found in [AS04].

Proof of Lemma 12. The proof uses the *probabilistic method*. To prove that there exists an object with a specific property, we find a way to select the object at random so that with positive probability the property is satisfied. In our case, the object is the collection of subsets $I_1, \ldots, I_\ell$. The properties are 20 and 21.

We first define a way to select these subsets at random. This is done with the following sampling procedure:

Algorithm 1 Sampling Procedure

for j := $1, \ldots, \ell$ **do**
  Draw a random set I'_j of coordinates (a subset of the set of all coordinates $[n]$), where each $i \in [n]$ is selected to be in I'_j independently with probability $\frac{\eta}{2}$.
  Set I_j to be the set of all coordinates $i \in I'_j$ such that $\sum_{k \in I'_j} |A_{ik}| \leq \eta$.
end for
Output $(I_1, \ldots, I_\ell)$

Notice that each subset is selected independently of the others. We are now going to prove that the output of this sampling procedure satisfies 20 and 21 with positive probability. By the nature of the procedure, 21 is automatically satisfied. It remains to check that 20 holds. First, for a fixed $i \in [n]$ and $j \in [\ell]$, we calculate the probability that $i \in I_j$.

$$\Pr\left[i \in I_j\right] = \Pr\left[i \in I'_j\right] \Pr\left[i \in I_j \middle| i \in I'_j\right] = \Pr\left[i \in I'_j\right] \Pr\left[\sum_{k \in I'_j} |A_{ik}| \leq \eta \;\middle|\; i \in I_j\right]$$

$$= \Pr\left[i \in I'_j\right] \Pr\left[\sum_{k \in I'_j} |A_{ik}| \leq \eta\right] = \frac{\eta}{2} \Pr\left[\sum_{k \in I'_j} |A_{ik}| \leq \eta\right].$$

By linearity of expectation we have:

$$\mathbb{E}\left[\sum_{k\in I'_j}|A_{ik}|\right] = \frac{\eta}{2}\sum_{k\in[n]}|A_{ik}| \leq \frac{\eta}{2}$$

Thus, by applying Markov's inequality we get:

$$\Pr\left[\sum_{k\in I'_j}|A_{ik}| \geq \eta\right] \leq \frac{1}{2},$$

which yields

$$\Pr\left[i \in I_j\right] \geq \frac{\eta}{4}.$$

Now, fix an $i \in [n]$ and define

$$S_i := |\{j \in [\ell] \colon i \in I_j\}|$$

Notice that the events $\{i \in I_j\}_{j=1}^{\ell}$ are independent from each other. Hence, we can write

$$S_i = \sum_{j=1}^{\ell} \mathbf{1}(i \in I_j)$$

which means that S_i is a sum of independent Bernoulli random variables. By the preceding calculation, we get:

$$\mathbb{E}[S_i] = \sum_{j=1}^{\ell}\mathbb{E}\left[\mathbf{1}(i \in I_j)\right] = \sum_{j=1}^{\ell}\Pr\left[i \in I_j\right] \geq \frac{\eta\ell}{4}.$$

Now, using Hoeffding's inequality and setting $\ell = 32\log 4\log n/\eta^2$ we get:

$$\begin{aligned}\Pr\left[S_i \leq \frac{\eta\ell}{8}\right] &\leq \Pr\left[|S_i - \mathbb{E}[S_i]| \geq \frac{\eta\ell}{8}\right] \leq 2\exp\left(-2\frac{\left(\frac{\eta\ell}{8}\right)^2}{\ell}\right)\\ &\leq 2\exp\left(-\frac{\eta^2\ell}{32}\right) \leq \frac{1}{2n}.\end{aligned}$$

By a simple union bound we get

$$\Pr\left[\exists i \in [n] : S_i \leq \frac{\eta\ell}{8}\right] \leq n\frac{1}{2n} = \frac{1}{2}$$

That means that with positive probability we have $S_i \geq \eta\ell/8$. Hence, there exists a collection of subsets $I_1, \ldots, I_\ell$ having this property. Notice that we do not have Equation 20 but an inequality instead. However, for each i such that $S_i > \eta\ell/8$, we can remove i from some sets so that the number of sets it appears is exactly $\lceil \eta\ell/8 \rceil$. Clearly, by removing elements from a set, inequality 21 still holds. □

Proof of Lemma 2. Let J be the interaction matrix of the Ising model distribution of X and let h denote the external field. By the hypothesis, we have that $\|J\|_\infty / M \leq 1$. Hence, by applying Lemma 12 for the matrix $\|J\|_\infty / M \leq 1$ with $\eta' = \eta / M$, we obtain that there is a collection of subsets $I_1, \ldots, I_\ell$ such that

- For all $i \in [n]$,
$$|\{j \in [\ell] \colon i \in I_j\}| = \lceil \frac{\eta \ell}{8M} \rceil$$
- For all $j \in [l]$ and all $i \in I_j$
$$\sum_{k \in I_j} \frac{|A_{ik}|}{M} \leq \frac{\eta}{M} \implies \sum_{k \in I_j} |A_{ik}| \leq \eta \tag{22}$$

We now argue that if inequality 22 holds, then the conditional distribution of X_{I_j} conditioned on X_{-I_j} is an (η, γ)-Ising model. To show that, it suffices to argue that $\Pr[X_{I_j} = y | X_{-I_j} = x_{-I_j}] \propto \exp(y^\top J' y + h'^\top y)$, for some interaction matrix J' and external field h'. Hence, we calculate the ratio of conditional probabilities for two configurations y and y':

$$\frac{\Pr\left[X_{I_j} = y | X_{-I_j} = x_{-I_j}\right]}{\Pr\left[X_{I_j} = y' | X_{-I_j} = x_{-I_j}\right]} = \frac{\exp\left(\sum_{u \in I_j, v \in I_j} J_{uv} y_u y_v + \sum_{u \in I_j, v \notin I_j} J_{uv} y_u x_v + h^\top y\right)}{\exp\left(\sum_{u \in I_j, v \in I_j} J_{uv} y'_u y'_v + \sum_{u \in I_j, v \notin I_j} J_{uv} y'_u x_v + h^\top y'\right)}$$

By the preceding equality, it is clear that the conditional distribution of X_{I_j} conditional on $X_{-I_j} = x_{-I_j}$ is an Ising model with interaction matrix $J' = \{J_{uv}\}_{u,v \in I_j}$ and external field $h'_i = h_i + \sum_{v \notin I_j} J_{iv} x_v$. This proves the claim.

To conclude with the last part of the lemma, fix $a \in \mathbb{R}^n$, and drawing $j \in [\ell]$ uniformly at random, one obtains

$$\mathbb{E}_j[\sum_{i \in I_j} a_i] = \mathbb{E}_j[\sum_{i=1}^n a_i \mathbf{1}(i \in I_j)] = \frac{\lceil \eta \ell / (8M) \rceil}{\ell} \sum_{i=1}^n a_i. \geq \frac{\eta}{8M} \sum_{i=1}^n a_i.$$

In particular, there exists some $j \in [\ell]$ which achieves a value of at least this expectation. □

9 Concentration Inequalities for General Functions

One of the main goals of the proof is to show that the derivative of the log pseudo-likelihood concentrates around its mean value. To do this, we rely on suitably conditioning on some of the spins, which guarantees that the conditional distribution on the remaining spins satisfies Dobrushin's condition. However, the task of showing concentration of the derivative in this regime remains. We begin with an overview of the results regarding concentration of functions on the boolean hypercube when Dobrushin's condition is satisfied. We then state and prove a modification of these results that serves our purposes well.

Many concentration results in the weak dependence regime concern functions that are multilinear polynomials. Specifically, in [Ada+19] they prove general concentration results for arbitrary multilinear polynomials of weakly dependent random variables. For polynomials of degree 2, they show the following:

Theorem 5 ([Ada+19]). *Let A be a symmetric matrix with zeros in the diagonal and X be an (α, γ)-bounded random variable. Let $p(x) = x^\top Ax$. Then, for any $t > 0$,*

$$\Pr\left[|p(X) - \mathbb{E}p(X)| > t\right] \leq \exp\left(-c \min\left(\frac{t^2}{\|A\|_F^2 + \|\mathbb{E}Ax\|_2^2}, \frac{t}{\|A\|_2}\right)\right).$$

where the constant c only depends on α, γ and not on the entries of A.

This inequality is tight up to constants. A nice property of this result is that it explicitly connects the radius of concentration with the Frobenius norm of the matrix.

Unfortunately, the derivative of the PMLE is not a polynomial. Hence, we cannot directly apply Theorem 5. Instead, we will modify it's proof so that it holds for arbitrary functions over the hypercube. The original proof relied on the fact that the Hessian matrix of a second degree polynomial is constant. Despite this not being true in our case, we will follow the same strategy and prove concentration using the second order Taylor approximation of the function. First, we need to define these quantities precisely. In the following, for a vector x and an index i, we denote by x_{i+} the vector obtained from x by replacing that i'th coordinate with 1 and by x_{i-} the one that is obtained by replacing this coordinate with -1.

Definition 4. *For an arbitrary function $f : \{0,1\}^n \mapsto \mathbb{R}$, we define the* discrete derivative *of the function as*

$$D_i f(x) := \frac{f(x_{i+}) - f(x_{i-})}{2}$$

Let $Df(x)$ denote the n-coordinate vector of discrete derivatives. Similarly, the function $H : \{0,1\}^n \mapsto \mathbb{R}^{n \times n}$ defined as

$$H_{ij}(x) = D_i(D_j f(x))$$

is called the discrete Hessian *of f.*

As we will see, our concentration bound will depend on the discrete derivative and Hessian of a function. However, in some cases it is more convenient to provide bounds for these quantities rather than explicitly calculate them. Therefore, we can replace these two quantities by the ones defined below.

Definition 5. *Let $f : \{0,1\}^n \mapsto \mathbb{R}$ be an arbitrary function. A function $\widetilde{D} : \{0,1\}^{n_1} \mapsto \mathbb{R}^n$ is called a* pseudo discrete derivative *for f if*

$$\|\widetilde{D}(x)\|_2 \geq \|Df(x)\|_2$$

for all $x \in \{0,1\}^n$. Additionally, we say that a function $\widetilde{H} \colon \{-1,1\}^n \to \mathbb{R}^{n_1 \times n_2}$ is a pseudo discrete Hessian *with respect to the pseudo discrete derivative $\widetilde{D}$ if for all $u \in \mathbb{R}^{n_1}, x \in \mathbb{R}^n$,*

$$\|u^\top \widetilde{H}(x)\|_2 \geq \|D(u^\top \widetilde{D}(x))\|_2.$$

We are now ready to state the modification of Theorem 5.

Theorem 6. *Let $f : \{0,1\}^n \mapsto \mathbb{R}$ be an arbitrary function and X an (α, γ)-bounded random variable. Then*

$$\Pr[|f(x) - \mathbb{E}f(x)| > t] \leq \exp\left(-c \min\left(\frac{t^2}{\|\mathbb{E}\widetilde{D}(x)\|_2^2 + \max_x \|\widetilde{H}(x)\|_F^2}, \frac{t}{\max_x \|\widetilde{H}(x)\|_2}\right)\right).$$

The proof of Theorem 6 follows the structure of Theorem 5, with only slight modifications. Hence, we are now going to describe the machinery used in the proof of Theorem 5.

The main ingredient of the proof is a discrete logarithmic Sobolev inequality, first proven in [Mar03]. First, we need a definition.

Definition 6. *For any function $f : \{0,1\}^n \mapsto \mathbb{R}$ and $i \in [n]$, we define*

$$\mathfrak{d}_i f(x) = \frac{1}{2}\left(\mathbb{E}\left(\left(f(X) - f(X_1, \dots, X_{i-1}, \widetilde{X}_i, X_{i+1}, \dots, X_n)\right)^2 \mid X = x\right)\right)^{1/2}$$

where the random variable $X = (X_1, \dots, X_n)$ has distribution μ and the conditional distribution of $\widetilde{X}_i$ given that $X = x$ is $\mu(\cdot|\bar{x}_i)$, which is the conditional distribution of X_i given that $X_j = x_j$ for $j \neq i$. We denote by $\mathfrak{d}f(x)$ the vector in $\mathbb{R}^n$ whose i-th coordinate is $\mathfrak{d}_i f(x)$.

This quantity intuitively captures how much function f changes on average when we resample one of its input variables independently. By the classical theory of Concentration Inequalities for i.i.d. random variables, we know that this *average Lipschitzness* directly affects the concentration properties of the function. This is indeed the case here, as the next lemma shows.

Lemma 13 ([Mar03; GSS19])**.** *Let $f : \{0,1\}^n \mapsto \mathbb{R}$ be an arbitrary function on the hypercube. Suppose X is an (α, γ)-bounded random variable. Then for any $p \geq 2$ it holds*

$$\|f(X) - \mathbb{E}f(X)\|_p := \left(\mathbb{E}\left(f(X) - \mathbb{E}f(X)\right)^p\right)^{1/p} \leq \sqrt{2Cp}\left(\mathbb{E}\left(\|\mathfrak{d}f(X)\|_2^p\right)\right)^{1/p}$$

where C is a function of α, γ which is bounded when α is bounded from 1 and γ is bounded from zero.

Lemma 13 essentially tells us that in order to bound the p-th moment of the function we wish to show concentration for, it is enough to control the p-th moment of $\|\mathfrak{d}f(X)\|_2$. In the proof of Theorem 5, the authors bound this moment by the corresponding moment of a multilinear form of gaussian random variables. In doing so, they exploit the fact that $\mathfrak{d}_i f(X)$ can be very conveniently bounded when f is a multilinear polynomial. We will follow exactly the same technique, while relying on the discrete derivative of f instead to bound $\mathfrak{d}_i f(X)$.

Proof of Theorem 6. As mentioned earlier, the proof follows the same strategy as [Ada+19], with a small adjustment. Our general strategy will be to bound the p-th moment of $f(X) - \mathbb{E}f(X)$ by the p-th moment of a gaussian multilinear form. By Jensen's Inequality, we have:

$$\begin{aligned}
\mathbb{E}\left(\|\mathfrak{d}f(X)\|_2^p\right) &= \mathbb{E}\left[\left(\sum_{i=1}^n \left(\frac{1}{2}\left(\mathbb{E}\left(\left(f(X) - f(X_1, \dots, X_{i-1}, \widetilde{X}_i, X_{i+1}, \dots, X_n)\right)^2 \mid X = x\right)\right)\right)\right)^{p/2}\right] \\
&\leq \mathbb{E}\left[\left(\sum_{i=1}^n \frac{1}{2}\left(f(X) - f(X_1, \dots, X_{i-1}, \widetilde{X}_i, X_{i+1}, \dots, X_n)\right)^2\right)^{p/2}\right]
\end{aligned}$$

We now note that for each i, either X_i and $\widetilde{X}_i$ will be the same or they will have opposite signs. This means that in any case

$$\left|f(X) - f(X_1, \dots, X_{i-1}, \widetilde{X}_i, X_{i+1}, \dots, X_n)\right| \leq |D_i f(X)|$$

This means that

$$\mathbb{E}\left(\|\eth f(X)\|_2^p\right) \leq \mathbb{E}\left(\|\frac{1}{2}Df(X)\|_2^p\right)$$

Combining this with the log Sobolev inequality of Lemma 13 we get:

$$\|f(X) - \mathbb{E}f(X)\|_p \leq \sqrt{Cp}\left(\mathbb{E}\|Df(X)\|_2^p\right)^{1/p} \leq \sqrt{Cp}\left(\mathbb{E}\|\widetilde{D}(X)\|_2^p\right)^{1/p} \tag{23}$$

Now suppose $g \sim \mathcal{N}(0, I_n)$ is an n-dimensional gaussian vector independent of X. For a fixed X, the random variable $\frac{\widetilde{D}(X)^\top g}{\|\widetilde{D}(X)\|_2}$ is a single dimensional gaussian $\mathcal{N}(0,1)$. Hence, by elementary properties of the gaussian distribution, there exists a constant $M > 0$ independent of p such that

$$\left(\mathbb{E}_g\left(\frac{\widetilde{D}(X)^\top g}{\|\widetilde{D}(X)\|_2}\right)^p\right)^{1/p} \geq \frac{1}{M}\sqrt{p}$$

The symbol $\mathbb{E}_g$ means that we are integrating only with respect to the random variable g. This implies that for a fixed X

$$\sqrt{p}\|\widetilde{D}(X)\|_2 \leq M\left(\mathbb{E}_g\left(\widetilde{D}(X)^\top g\right)^p\right)^{1/p}$$

Combining with Equation 23 , we get that for all functions f and pseudo derivatives $\widetilde{D}$:

$$\|f(X) - \mathbb{E}f(X)\|_p \leq K\left(\mathbb{E}_{X,g}\left(\widetilde{D}(X)^\top g\right)^p\right)^{1/p} \tag{24}$$

where $K = M\sqrt{C}$. Inequality 24 is a first step in proving the bound on the p-th moment. We want to use this inequality to make the second derivative appear on the right hand side. To do this, we "fix" the value of g. By Minkowski's inequality we have:

$$\left(\mathbb{E}_{X,g}\left(\widetilde{D}(X)^\top g\right)^p\right)^{1/p} = \|\widetilde{D}(X)^\top g\|_p \leq \|\widetilde{D}(X)^\top g - \mathbb{E}_X\widetilde{D}(X)^\top g\|_p + \|\mathbb{E}_X\widetilde{D}(X)^\top g\|_p$$

First, we bound the second term. By linearity of expectation, we have:

$$\|\mathbb{E}_X\widetilde{D}(X)^\top g\|_p = \left(\mathbb{E}_g\left((\mathbb{E}_X\widetilde{D}(X))^\top g\right)^p\right)^{1/p}$$

Now, the variable $(\mathbb{E}_X\widetilde{D}(X))^\top g$ is clearly a single dimensional gaussian, which means that it's p-th moment is bounded as:

$$\|\mathbb{E}_X\widetilde{D}(X)^\top g\|_p \leq M\sqrt{p}\|\mathbb{E}_X\widetilde{D}(X)\|_2 \tag{25}$$

and that concludes the bound for the second term. For the first term, fix g and define the function $h : \{0,1\}^n \mapsto \mathbb{R}$ as $h(x) = \widetilde{D}(x)^\top g - \mathbb{E}_X\widetilde{D}(X)^\top g$. Now, we apply inequality 23 to the function h, which gives us

$$\begin{aligned}
\left(\mathbb{E}_X\left(\widetilde{D}(X)^\top g - \mathbb{E}_X\widetilde{D}(X)^\top g\right)^p\right)^{1/p} &\leq \sqrt{Cp}\left(\mathbb{E}_X\|D(\widetilde{D}(X)^\top g)\|_2^p\right)^{1/p} \\
&\leq \sqrt{Cp}\left(\mathbb{E}_X\|g^\top\widetilde{H}(X)\|_2^p\right)^{1/p} \\
&\leq \sqrt{C}M\left(\mathbb{E}_{X,g'}\left|g^\top\widetilde{H}(X)g'\right|^p\right)^{1/p}
\end{aligned}$$

We conclude that

$$\begin{aligned}\|\widetilde{D}(X)^\top g - \mathbb{E}_X \widetilde{D}(X)^\top g\|_p &= \left(\mathbb{E}_{X,g}\left(\widetilde{D}(X)^\top g - \mathbb{E}_X \widetilde{D}(X)^\top g\right)^p\right)^{1/p} \\ &\leq K\left(\mathbb{E}_{X,g,g'}\left|g^\top \widetilde{H}(X) g'\right|^p\right)^{1/p}\end{aligned}$$

Notice now that we have to bound the p-th norm of a quadratic form $g^\top \widetilde{H}(X) g'$ where g, g' are independent gaussian vectors. That is precisely what the Hanson-Wright inequality gives us. Note that the matrix $\widetilde{H}$ need not have zeroes in the diagonal, as mentioned in [RV13; Lat+06]. [5] For a fixed X there is a constant C' such that:

$$\left(\mathbb{E}_{g,g'}\left(g^\top \widetilde{H}(X) g'\right)^p\right)^{1/p} \leq C'\left(\sqrt{p}\|\widetilde{H}(X)\|_F + p\|\widetilde{H}(X)\|_2\right) \leq C'\left(\sqrt{p}\max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_F + p\max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_2\right)$$

This gives us the final inequality:

$$\|\widetilde{D}(X)^\top g - \mathbb{E}_X \widetilde{D}(X)^\top g\|_p \leq C'\left(\sqrt{p}\max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_F + p\max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_2\right) \tag{26}$$

Putting together inequalities 25 and 26 we conclude that there exists a universal constant C'' such that

$$\|f(X) - \mathbb{E}f(X)\|_p \leq C''\left(\sqrt{p}\|\mathbb{E}_X \widetilde{D}(X)\|_2 + \sqrt{p}\max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_F + p\max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_2\right)$$

To conclude the proof, we notice that by Markov's inequality and the preceding result:

$$\Pr\left[|f(X) - \mathbb{E}f(X)| > eC''\left(\sqrt{p}\|\mathbb{E}_X \widetilde{D}(X)\|_2 + \sqrt{p}\max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_F + p\max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_2\right)\right] \leq e^{-p}$$

We now set

$$p = \min\left(\frac{t^2}{\|\mathbb{E}_X \widetilde{D}(X)\|_2^2 + \max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_F^2}, \frac{t}{\max_{x\in\{0,1\}^n}\|\widetilde{H}(x)\|_2}\right)$$

and the result follows. □

10 Improved Bound for Estimating a Single Parameter

In this section, we prove Corollary 10, which we now restate for convenience.

Corollary 10. *Let $M > 0$, let J_0 be a fixed matrix with $\|J_0\|_\infty \leq 1$ and let β^* be some unknown parameter satisfying $|\beta^*| \leq M$. Then, there exists an estimator $\hat{\beta}$ from a single sample $x \sim \Pr_{\beta^* J_0}$ such that w.p. $\geq 1-\delta$, $|\hat{\beta} - \beta^*| \leq C(M)F(\beta^* J_0)^{-1/2}(\log\log n + \log(1/\delta))$ where $F(\cdot)$ is defined as in* (1).

[5] The standard result in [RV13; Lat+06] applies to square matrices A. Since $\widetilde{H}$ might not be necessarily a square matrix, we can just make it square by adding zeroes to the missing dimensions. The quadratic form doesn't change and the matrix norms $\|\cdot\|_F, \|\cdot\|_2$ remain the same.

Similarly to the proof of Theorem 4, our estimator is the maximum pseudo likelihood. Define the negative log pseudo-likelihood as

$$\varphi(\beta) = -\sum_{i=1}^{n} \log \Pr_{\beta J}[x_i \mid x_{-i}].$$

Then, the following holds:

$$\varphi'(\beta) := \frac{d\varphi(\beta)}{d\beta} = \frac{1}{2}\sum_{i=1}^{n}(J_i x)(\tanh(\beta J_i x) - x_i) \tag{27}$$

and further,

$$\varphi''(\beta) := \frac{d^2\varphi(\beta)}{d\beta^2} = \frac{1}{2}\sum_{i=1}^{n}(J_i x)^2 \operatorname{sech}^2(\beta J_i x) \geq c\|Jx\|_2^2, \tag{28}$$

where the last inequality holds uniformly for all $|\beta| \leq M$, since sech^2 is lower bounded whenever its argument is bounded from above. Similarly to the arguments in Lemma 4, it suffices to bound the ratio between the first and second derivatives of φ and we obtain that

$$|\beta^* - \hat{\beta}| \leq \frac{|\varphi'(\beta^*)|}{\min_{|\beta|\leq M}\varphi''(\beta^*)} \leq C\frac{|\varphi'(\beta^*)|}{\|Jx\|_2^2}. \tag{29}$$

We prove a lower bound on $\|Jx\|_2^2$ based on arguments from [Cha07; BM18], and use lemmas from the proof of Theorem 4 to show that the derivative is bounded in terms of $\|Jx\|_2$.

We begin with the second part. For this purpose, we provide a restatement of the lemmas from the proof of Theorem 4 that we will use here. First, the sub-sampling lemma, that shows how to reduce the correlations in the Ising model by subsampling:

Lemma 2. *Let $x = (x_1, \ldots, x_n)$ be an (M, γ)-Ising model, and fix $\eta \in (0, M]$. Then, there exist subsets $I_1, \ldots, I_\ell \subseteq [n]$ with $\ell \leq CM^2 \log n/\eta^2$ such that:*

1. *For all $i \in [n]$,*
$$|\{j \in [\ell] \colon i \in I_j\}| = \left\lceil \frac{\eta\ell}{8M} \right\rceil$$
.
2. *For all $j \in [\ell]$ and any value of x_{-I_j}, the conditional distribution of x_{I_j} conditioned on x_{-I_j} is an (η, γ)-Ising model.*

Furthermore, for any non-negative vector $\theta \in \mathbb{R}^n$ there exists $j \in [\ell]$ such that

$$\sum_{i\in I_j}\theta_i \geq \frac{\eta}{8M}\sum_{i=1}^{n}\theta_i.$$

We create the sets $I_1, \ldots, I_\ell$ from this lemma with $\eta = 1/2$. The following two lemmas would provide an upper bound on φ' and a lower bound on $\|Jx\|_2$, respectively, both in terms of the same quantity, which is a function of the sets $I_1, \ldots, I_\ell$:

Lemma 14 (A special case of Lemma 3)**.** *For any $t > 0$,*

$$\Pr\left[\left|\varphi'(\beta^*)\right| > t\left(\|J\|_F + \max_j \|\mathbb{E}[Jx|x_{-I_j}]\|_2\right)\right] \le C\log n \exp\left(-c\min\left(t^2, \frac{t\|J\|_F}{\|J\|_2}\right)\right).$$

Lemma 15 (A special case of Lemma 4)**.** *For any $t > 0$:*

$$\Pr\left[\|Jx\|_2^2 < c\|J\|_F^2 + c\max_j \|\mathbb{E}[Jx \mid x_{-I_j}]\|_2^2 - t\left(\|J\|_F + \max_j \|\mathbb{E}[Jx \mid x_{-I_j}]\|_2\right)\right]$$
$$\le C\log n \exp\left(-c\frac{\min(t^2, t\|J\|_F)}{\|J\|_2^2}\right).$$

Lemmas 14 and 15 follow from Lemma 3 and Lemma 4 by substituting J^* with $\beta^* J$ and A with J, and noting that the quantity $\frac{\partial\varphi(J\beta^*)}{\partial J}$ in Section 6 corresponds to $\varphi'(\beta^*)$ in this section. As a direct corollary, we can bound φ' with respect to $\|Jx\|$:

Lemma 16. *For any $t \ge 1$, with probability at least $1 - \log n e^{-ct}$,*

$$\left|2\varphi'(\beta^*)\right| = \left|x^\top Jx - \sum_{i=1}^n J_i x \tanh(\beta^* J_i x)\right| \le Ct\|Jx\|_2 + Ct^2.$$

Proof. The first equality follows from definition, hence we will prove the inequality. From Lemma 14, it holds that for any $t \ge 1$,

$$\Pr\left[\left|\varphi'(\beta^*)\right| > t\left(\|J\|_F + \max_j \|\mathbb{E}[Jx|x_{-I_j}]\|_2\right)\right] \tag{30}$$
$$\le C\log n \exp\left(-c\min\left(t^2, \frac{t\|J\|_F}{\|J\|_2}\right)\right) \tag{31}$$
$$\le C\log n \exp(-c\min(t^2, t)) \tag{32}$$
$$= C\log n \exp(-ct), \tag{33}$$

since $\|J\|_F \ge \|J\|_2$ for all J and since $t^2 \ge t$ for all $t \ge 1$. Second of all, from Lemma 15, for any $t \ge 1$, we have that

$$\Pr\left[\|Jx\|_2^2 < c\|J\|_F^2 + c\max_j \|\mathbb{E}[Jx \mid x_{-I_j}]\|_2^2 - t\left(\|J\|_F + \max_j \|\mathbb{E}[Jx \mid x_{-I_j}]\|_2\right)\right] \tag{34}$$
$$\le C\log n \exp\left(-c\frac{\min(t^2, t\|J\|_F)}{\|J\|_2^2}\right) \tag{35}$$
$$\le C\log n \exp(-c\min(t^2, t)) \tag{36}$$
$$\le C\log n \exp(-ct). \tag{37}$$

using the fact that $\|J\|_F \ge \|J\|_2$ and that $\|J\|_2$ is at most some constant.

Next, with probability at least $1 - 2\log n e^{-ct}$, both the events that their probabilities are estimated in (30) and (34) hold, and assume that they do hold till the rest of the proof. Let $\zeta = \|J\|_F + \max_j \|\mathbb{E}[Jx \mid x_{-I_j}]\|_2$ and we derive that $|\varphi'(\beta^*)| \le t\zeta$ and that $\|Jx\|_2^2 \ge c\zeta^2 - t\zeta$. It follows that

$$\left|\varphi'(\beta^*)\right| \le Ct\|Jx\|_2 + Ct^2. \tag{38}$$

Indeed, if $t \le c\zeta/2$ then

$$\|Jx\|_2^2 \ge c\zeta^2/2 \ge c(|\varphi'(\beta^*)|/t)^2/2$$

hence

$$|\varphi'(\beta^*)| \le \sqrt{2/c}t\|Jx\|_2.$$

Otherwise,

$$|\varphi'(\beta^*)| \le t\zeta \le 2t^2/c.$$

This concludes the proof. □

Next, we lower bound $\|Jx\|_2^2$ in terms of the log partition function $F(\beta^* J)$, based on arguments from [Cha07; BM18].

$$\|Jx\|_2^2 = \sum_{i=1}^n (J_i x)^2 \ge \sum_{i=1}^n \frac{J_i x \tanh(\beta^* J_i x)}{\beta^*}, \tag{39}$$

since for all $y \in \mathbb{R}$, y and $\tanh(y)$ have the same sign and additionally, $|\tanh(y)| \le |y|$. From Lemma 16, the right hand side of (39) can be approximated by $x^\top Jx/\beta^*$. Further, the term $x^\top Jx$ can be lower bounded by the log partition function:

Lemma 17. *With probability at least $1 - e^{-F(J\beta^*)/2}$, $x^\top Jx \ge F(J\beta^*)/\beta^*$.*

Proof. Recall that

$$\Pr_{J\beta^*}[x] = 2^{-n} e^{\beta^* x^\top Jx/2 - F(J\beta^*)}$$

hence,

$$\mathbb{E}_{J\beta^*}\left[e^{-\beta^* x^\top Jx/2}\right] = \sum_x e^{-\beta^* x^\top Jx/2} \Pr_{J\beta^*}[x] = e^{-F(J\beta^*)}.$$

Therefore, by Markov's inequality,

$$\begin{aligned}\Pr_{J\beta^*}\left[\beta^* x^\top Jx < F(J\beta^*)\right] &= \Pr_{J\beta^*}\left[e^{-\beta^* x^\top Jx/2} > e^{-F(J\beta^*)/2}\right] \\ &\le \mathbb{E}_{J\beta^*}[e^{-\beta^* x^\top Jx/2}]/e^{-F(J\beta^*)/2} = e^{-F(J\beta^*)/2}.\end{aligned}$$

□

By the above lemmas, we derive the following:

Lemma 18. *For any $1 \le t \le c\sqrt{F(J\beta^*)}$, with probability at least e^{-ct},*

$$\frac{\varphi'(\beta^*)}{\|Jx\|_2^2} \le Ct\beta^*/\sqrt{F(J\beta^*)}.$$

Proof. We derive from (39), Lemma 16 and Lemma 18, and from $t \le c\sqrt{F(\beta^* J)}$, that with probability at least $\log ne^{-ct}$,

$$\begin{aligned}\|Jx\|_2^2 \ge \frac{1}{\beta^*}\sum_{i=1}^n J_i x \tanh(\beta^* J_i x) &\ge \frac{x^\top Jx - Ct\|Jx\|_2 - Ct^2}{\beta^*} \ge \frac{F(J\beta^*) - Ct^2\beta^*}{(\beta^*)^2} - \frac{Ct\|Jx\|_2}{\beta^*} \\ &\ge \frac{F(J\beta^*)/2}{(\beta^*)^2} - \frac{Ct\|Jx\|_2}{\beta^*},\end{aligned} \tag{40}$$

where the last inequality follows from the above assumption that $1 \leq t \leq c\sqrt{F(J\beta^*)}$ for a sufficiently small $c > 0$ and the $|\beta|$ is bounded by a constant. Assume that (40) holds from now onward. We derive that

$$(\|Jx\|_2\beta^*)^2 + \|Jx\|_2\beta^* Ct - F(J\beta^*)/2 \geq 0.$$

By solving the quadratic inequality for $\|Jx\|_2\beta^*$, we derive that

$$\|Jx\|_2\beta^* \geq \frac{-Ct + \sqrt{C^2t^2 + 2F(J\beta^*)}}{2} \geq \sqrt{F(J\beta^*)/2} - Ct.$$

Applying the bound $t \leq c\sqrt{F(J\beta^*)}$, we derive that

$$\|Jx\|_2 \geq \sqrt{F(J\beta^*)}/(2\beta^*). \tag{41}$$

Further, from Lemma 16, with probability at least $\log ne^{-ct}$,

$$\varphi'(\beta^*) \leq Ct\|Jx\|_2 + Ct, \tag{42}$$

and assume that this holds for the rest of the proof. By (41) and (42),

$$\frac{\varphi'(\beta^*)}{\|Jx\|_2^2} \leq C\frac{t\|Jx\|_2 + t}{\|Jx\|_2^2} = \frac{Ct}{\|Jx\|_2} + \frac{Ct}{\|Jx\|_2^2} \leq \frac{2\beta^* Ct}{\sqrt{F(J\beta^*)}} + \frac{4(\beta^*)^2 Ct}{F(J\beta^*)} \leq \frac{C't\beta^*}{\sqrt{F(J\beta^*)}}, \tag{43}$$

by the assumption of this lemma that $\sqrt{F(J\beta^*)}$ is at least a constant and since β^* is bounded by a constant. □

By (29) and by Lemma 18, Corollary 10 follows, by taking $t = \log\log n + \log(1/\delta)$.

11 The Lower Bound on $\|\hat{J} - J^*\|_F$

The estimation error of Theorem 4 is optimal in many interesting applications. In this Section, we present the proof of the general lower bound on the estimation rate of $\|\hat{J} - J^*\|$ that only depends on the packing numbers of the set $\mathcal{J}$. It is well known that the packing numbers are very related to the covering numbers of a set. We note that the expression for the lower bound does not match the one from the upper bound. However, this is to be expected, since for arbitrary sets $\mathcal{J}$ we might need more information other than the covering number to characterize learnability. However, in the particular case where $\mathcal{J}$ is a linear k-dimensional subspace of matrices, the lower bound becomes (nearly)-optimal up to constants when $\|J^*\| < 1$. This implies that doing MPLE in high temperature is optimal. We first present the general tool for proving our lower bounds, which is going to be Fano's method. The following result is adapted from [Duc16].

Lemma 19 ([Km79], Fano's method)**.** *Let $\mathcal{P} = \{P_J \colon J \in \mathcal{J}\}$ be a family of distributions indexed by matrices $J \in \mathcal{J}$. For a distribution P from this family, denote $J(P)$ the interaction matrix corresponding to it. Let $\{P_v\}_{v\in\mathcal{V}}$ be a finite subset of distributions from $\mathcal{P}$ such that for any $v \neq v' \in \mathcal{V}$ we have $\|J(P_v) - J(P_{v'})\|_F \geq 2r$ for some $r > 0$ (this family is also called a $2r$-packing for $\|\cdot\|_F$). Assume that V is a random variable that is uniform on the set $\mathcal{V}$, and conditional on $V = v$, and we draw a sample $X \sim P_v$. Then the miminax risk for estimation with respect to the Frobenius norm is lower bounded by*

$$r\left(1 - \frac{I(V;X) + \log 2}{\log|\mathcal{V}|}\right)$$

A standard way of using Fano's method is finding a $r > 0$ and a family $\{P_v\}_{v \in \mathcal{V}}$ such that

$$\frac{I(V;X) + \log 2}{\log |\mathcal{V}|} \leq \frac{1}{2}$$

This would immediately imply the lower bound $r/2$. To prove this statement, we need to upper bound the mutual information $I(V;X)$. If the mutual information is small, then this means that by seeing the sample X we cannot infer with large probability the distribution V of the family that generated it. To bound this quantity, we will use the following inequality, whose simple proof can be found in Chapter 7 of [Duc16].

$$I(V;X) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v,v'} D_{KL}(P_v \| P_{v'}) \tag{44}$$

Thus, if we can bound the KL-divergence between all pairs in the family, we also have the same bound for $I(V;X)$. The following Lemma aims to provide such a bound, when the two models are in high temperature.

Lemma 20. *Suppose P_{J_1}, P_{J_2} be the distributions of two Ising models with interaction matrices J_1 and J_2, respectively. Suppose furthermore that $\|J_1\|_\infty, \|J_2\|_\infty < 1 - \alpha$ for some $\alpha \in (0,1)$. Then, there is a constant $C = C(\alpha)$ such that*

$$D_{KL}(P_{J_1} \| P_{J_2}) \leq C \|J_2 - J_1\|_F^2$$

Proof. In the following computations, the concept of the log partition function will be useful. We thus define

$$F(J) = \ln \sum_{y \in \{-1,+1\}^n} \exp\left(\frac{1}{2} y^\top J y\right)$$

which is the log partition function of a model with interaction matrix J. We also define the following function $g : [0, \|J_2 - J_1\|_F] \to \mathbb{R}$ by

$$g(t) = F\left(J_1 + tS\right), \quad \text{where } S = \frac{J_2 - J_1}{\|J_2 - J_1\|_F}.$$

First, we notice that $g(0) = F(J_1), g(\|J_2 - J_1\|_F) = F(J_2)$. Also, some simple calculations show that

$$g'(t) = \mathbb{E}_{y \sim J_1 + tS}\left[\frac{1}{2} y^\top S y\right]$$
$$g''(t) = \mathrm{Var}_{y \sim J_1 + tS}\left[\frac{1}{2} y^\top S y\right]$$

where the notation $y \sim A$ means that y is sampled from an Ising model with interaction matrix

A. Now, we do some simple calculations with the KL-divergence

$$\begin{aligned} D_{KL}(P_{J_1}\|P_{J_2}) &= \mathbb{E}_{y\sim P_{J_1}} \ln \frac{P_{J_1}(y)}{P_{J_2}(y)} \\ &= \mathbb{E}_{y\sim P_{J_1}}\left[\frac{1}{2}y^\top(J_1-J_2)y - F(J_1) + F(J_2)\right] \\ &= F(J_2) - F(J_1) - \|J_2-J_1\|_F \mathbb{E}_{y\sim J_1}\left[\frac{y^\top S y}{2}\right] \\ &= g(\|J_2-J_1\|_F) - g(0) - \|J_2-J_1\|_F g'(0) \\ &= \frac{\|J_2-J_1\|_F^2}{2} g''(\xi) \end{aligned}$$

for some $\xi \in [0, \|J_2-J_1\|_F]$. In the last step, we used Taylor's Theorem on g. We have

$$\frac{\|J_2-J_1\|_F^2}{2} g''(\xi) = \frac{\|J_2-J_1\|_F^2}{2} \operatorname{Var}_{y\sim J_1+\xi S}\left[\frac{1}{2}y^\top S y\right] = \frac{1}{2}\operatorname{Var}_{y\sim J_1+\xi S}\left[\frac{1}{2}y^\top(J_2-J_1)y\right] \ .$$

Hence, it suffice to bound this variance in order to obtain a bound on the KL-divergence. This is the variance of a second degree polynomial of an Ising model with interaction matrix $J_1+\xi S$. It is easy to see that as t varies in $[0, \|J_2-J_1\|_F]$, J_1+tS moves along the line segment connecting J_1, J_2. Hence, $J_1+\xi S$ also belongs in this line segment. Since $\|J_1\|_\infty, \|J_2\|_\infty < 1-\alpha$, it follows that any point in the line segment connecting them has infinity norm upper bounded by $1-\alpha$. We conclude that $\|J_1+\xi S\|_\infty$ defines an Ising model in high temperature. The task is thus to bound a second degree polynomial of an Ising model in high temperature. Using the bound of Theorem 2.1 in [GLP+18] we obtain that there exists a constant $C = C(\alpha)$ such that

$$\operatorname{Var}_{y\sim J_1+\xi S}\left[\frac{1}{2}y^\top(J_2-J_1)y\right] \le C(\alpha)\|J_2-J_1\|_F^2 \ .$$

This implies the result of this Lemma. □

Now that we have a bound for KL-divergence between two Ising models, we can prove our general lower bound.

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[6] 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 $\max_{J^*\in\mathcal{J}} \mathbb{E}_{x\sim P_{J^*}}[\|\hat{J}(x)-J^*\|_F] \ge r/2$.

Proof. The proof is essentially an application of the Fano method when we use the upper bound on the KL proven in Lemma 20. Consider any $r > 0$ that satisfies the requirements of the Theorem. Then, by definition, there exists an $R > 0$, such that the distance of any two matrices in $\mathcal{J}$ is bounded by R in Frobenius norm. We need to define the following concept.

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

Definition 7 ([Ver18]). *A subset L of a metric space (T, d, ϵ) is* ϵ-separated *if $d(x, y) > \epsilon$ for any distinct $x, y \in L$. The largest possible cardinality of an ϵ separated subset of a given set $L \subseteq T$ is called the* packing number *of L and denoted by $\mathcal{P}(L, d, \epsilon)$*

The packing number of a set is the maximum amount of points that we can choose so that all of them are far from each other. Notice that this concept is very related to the requirements of Fano method. Indeed, to apply it we need a set of distributions that are far from each other.

To use these concepts, let $\mathcal{V}$ be a maximum cardinality $2r$ separated subset of $\mathcal{J}$ with respect to the Frobenius norm. This means that

$$|\mathcal{V}| = \mathcal{P}(\mathcal{J}, \|\cdot\|_F, 2r)$$

Then, if X, V are defined as in Lemma 19, we have that

$$I(V; X) \leq \frac{1}{|\mathcal{V}|^2} \sum_{v, v' \in \mathcal{V}} D_{KL}(P_v \| P_{v'}) \leq C(\alpha) R^2$$

which follows from Eq. (44) and Lemma 20 since $\mathcal{V} \subseteq \mathcal{J}$ and $\mathcal{J}$ has diameter at most R in Frobenius norm.

Now, by the assumption of this Theorem, we conclude that

$$I(V; X) + \log 2 \leq C(\alpha) R^2 + \log 2 \leq \frac{\log \mathcal{N}(\mathcal{J}, \|\cdot\|_F, 2r)}{2}.$$

We would like to have $|\mathcal{V}|$ appear on the right hand side. The covering and packing numbers of a set are connected by the following simple inequality, whose proof can be found in Chapter 4 of [Ver18].

$$\mathcal{N}(\mathcal{J}, \|\cdot\|_F, 2r) \leq \mathcal{P}(\mathcal{J}, \|\cdot\|_F, 2r)$$

Hence, we conclude that

$$I(V; X) + \log 2 \leq \frac{\log \mathcal{P}(\mathcal{J}, \|\cdot\|_F, 2r)}{2} = \frac{\log |\mathcal{V}|}{2}$$

Hence, we have that

$$\frac{I(V; X) + \log 2}{\log |\mathcal{V}|} \leq \frac{1}{2}$$

Thus, by Lemma 19 we conclude that the minimax risk is at least $r/2$, which concludes the proof. □

Theorem 2 offers a lower bound in terms of the covering numbers. However, the exact dependence of the minimax risk on the covering numbers is not clear from this general formulation. We now offer a simple Corollary of this theorem that gives (nearly)-optimal lower bounds when $\mathcal{J}$ is a linear subspace of matrices with bounded infinity norm.

Corollary 8. *Let $k \in \mathbb{N}$, let $J_1, \ldots, J_k$ be interaction matrices with disjoint supports*[7] *such that $\|J_i\|_\infty \leq 1$ and $\|J_i\|_F \geq k$ for all i. Define $\mathcal{J} = \{J = \sum_i \alpha_i J_i \colon \alpha_i \in \mathbb{R}, \|J\|_\infty \leq 1\}$. Then, any one-sample estimator $\hat{J}(x)$ has a minimax error of $\sup_{J^* \in \mathcal{J}} \mathbb{E}_{x \sim \Pr_{J^*}} \left[\|\hat{J} - J^*\|_F\right] \geq c\sqrt{k}$.*

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

Proof. Let $r = c_1\sqrt{k}$, where c_1 will be determined later. We will prove that r satisfies all the requirements of Theorem 2 for some suitable constant c_1. Suppose that we renormalize every matrix J_i as follows

$$J_i' = \frac{1}{2\|J_i\|_F} J_i$$

We know that $\|J_i\|_F \geq k$ for all i. Thus, for any set of $\alpha_i \in [-1, 1]$, we have that

$$\left\| \sum_{i=1}^{k} \alpha_i J_i' \right\|_\infty \leq k \frac{1}{2k} = \frac{1}{2}$$

hence $\sum_{i=1}^{k} \alpha_i J_i' \in \mathcal{J}$. Also

$$\|J_i'\|_F = \frac{1}{2}.$$

Now consider the family

$$\mathcal{V} = \left\{ \sum_{i=1}^{k} \sigma_i J_i' : \sigma_i \in \{-1, 1\} \right\}$$

We will prove that the minimax risk for estimation in this family is lower bounded. Since $\mathcal{V} \subseteq \mathcal{J}$, this would immediately imply a lower bound for the estimation rate on $\mathcal{J}$. Since the supports of J_i are disjoint, these matrices are orthogonal with respect to the trace inner product. This means that the maximum distance R between any pair of matrices in $\mathcal{V}$ is at most $2\sqrt{k}$. We will show that we can find a large number of matrices in $\mathcal{V}$ which are all at distance at least r from each other. This is equivalent to finding a lower bound for the packing number of $\mathcal{V}$.

To do so, we take a packing of the hypercube $\{-1, 1\}^k$ with respect to the Hamming distance. Namely, we take a set $U \subseteq \{-1, 1\}^k$ such that for any $\sigma, \sigma' \in U$, the Hamming distance between σ and σ', $\|\sigma - \sigma'\|_1/2$, is at least $k/3$. It is known that there exists such a set $\mathcal{U}$ with cardinality e^{ck} for some universal constant $c > 0$.

Given the set $\mathcal{U}$, we take the following set $\mathcal{V}'$ to be our r-packing of $\mathcal{V}$:

$$\mathcal{V}' = \left\{ \sum_{i=1}^{k} \sigma_i J_i' \colon \sigma \in \mathcal{U} \right\}.$$

Since the matrices J_i' have disjoint support, they are orthonormal with respect to the Forbenious norm, hence for any $\sigma \neq \sigma' \in \mathcal{U}$,

$$\left\| \sum_{i=1}^{k} \sigma_i J_i' - \sum_{i=1}^{k} \sigma_i' J_i' \right\|_F^2 = \sum_{i=1}^{k} \left\| (\sigma_i - \sigma_i')^2 J_i' \right\|_F^2 = \frac{1}{4} \sum_{i=1}^{k} (\sigma_i - \sigma_i')^2 \geq \frac{k}{3},$$

using the fact that each J_i' has Forbenius norm of $1/2$, and that $\sigma, \sigma' \in \mathcal{U}$, hence they differ in at least $k/3$ coordinates. Hence, by setting $r = \sqrt{k/3}/2$, we just found a $2r$ packing on the set $\mathcal{V}$ that contains at least e^{ck} elements. Hence,

$$\log \mathcal{P}(\mathcal{V}, \|\cdot\|_F, 2r) \geq ck$$

for some c. Now notice that in order to apply Theorem 2, the inequality that we have to satisfy is

$$C(\alpha) R^2 + \log 2 \leq \frac{\log \mathcal{P}(\mathcal{V}, \|\cdot\|_F, 2r)}{2}, \tag{45}$$

where $R = \max_{J \neq J' \in \mathcal{V}} \|J - J'\|_F$ (in fact, the statement of Theorem 2 requires a similar inequality with $\mathcal{P}$ replaced by $\mathcal{N}$, but it follows from the proof that Eq. (45) suffices). We essentially want to reduce R to a suitable multiple of $\sqrt{k}$ so that this inequality will hold. To do that, we simply scale down the J_i' by a suitable constant. Of course, we scale down r by the same constant. Notice that the packing numbers will not be affected by this, since we just scaled the whole space. Hence, the right hand side remains the same, while the left hand side is properly reduced. Hence, this inequality will hold for $r = c_1\sqrt{k}$ for some suitable c_1. This means by Theorem 2 that the minimax risk of estimation in Frobenius norm is $\Omega(\sqrt{k})$. □

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), and by the DARPA award HR00111990021.

A Algorithm for Maximizing Pseudo-Likelihood

In this Section, we show that there exists a polynomial time algorithm for the MPLE, which is the content of Lemma 21. The central problem is maximizing the pseudo-likelihood with the constraint that the matrix we output has low infinity norm. To achieve this, we use a regularizer of the form $\lambda \max(0, \|A_{\boldsymbol{\beta}}\|_\infty - M)$. This ensures that the optimal solution that the algorithm outputs will have small infinity norm. We should also argue that the output of the regularized procedure is close to the minimizer of φ. To do this, we need to calculate bounds for φ and its derivative. Since the regularized part is not differentiable, we have to use subgradient optimization methods.

Lemma 21. *There exists an algorithm that outputs $\hat{J}$ satisfying $\hat{J} \in V$ and $\|\hat{J}\|_\infty \leq 2M$, such that*

$$\varphi(\hat{J}) \leq \varphi(J^*) + \epsilon$$

and runs in time $O(k^2 n^6 M^2/\epsilon^2)$.

Proof. The algorithm essentially solves the constrained optimization problem described in 10. To design the optimization algorithm, it will be more convenient to work with a specific base of matrices that span $\mathcal{V}$. For this reason, let $\{A_1, \ldots, A_k\}$ be a fixed orthonormal basis of $\mathcal{V}$ with respect to the inner product induced by the frobenius norm. Then, each $J \in \mathcal{V}$ can be uniquely written in the form

$$J = \sum_{i=1}^{k} \beta_i A_i$$

We use the notation $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ and $A_{\boldsymbol{\beta}} = \sum_{i=1}^k \beta_i A_i$. Thus, minimizing $\varphi(J)$ subject to $J \in \mathcal{J}$ is the same as minimizing

$$\psi(\boldsymbol{\beta}) = \varphi(A_{\boldsymbol{\beta}})$$

Algorithm 2 Optimization Procedure

Input:
Basis of matrices $J_1, \ldots, J_k$
Accuracy ϵ
Sample $(x_1, \ldots, x_n)$
Infinity norm bound M

Output:
Vector $\hat{\beta}$

$(A_1, \ldots, A_k) :=$ GRAM-SCHMIDT$(J_1, \ldots, J_k)$// Find orthonormal basis
for i $:= 1, \ldots, k$ **do**
  $\beta_i^0 := 0$ // Initializations
end for
$T = \frac{M^2 n^4 k}{\epsilon^2}$ // Number of steps of gradient descent
$\eta := \frac{M}{n\sqrt{k}\sqrt{T}}$ // stepsize
for t $:= 1, \ldots, T$ **do**
  $U := \beta_1^t A_1 + \ldots \beta_k^t A_k$ // The interaction matrix at time t
  for i $:= 1, \ldots, k$ **do**
    $S_i := 0$
    for j $:= 1, \ldots, n$ **do**
      $S_i := S_i + ((A_i)_j x)\left(\tanh(U_j x) - x_j\right)$ // Compute the derivative
    end for
  end for
  if $\|U\|_\infty \geq M$ **then**
    $MAX := -\infty$ // If regularized part is nonzero, also need subgradient
    $ARGMAX := 1$ // First find max row sum
    for l $:= 1, \ldots, n$ **do**
      $TEMP := 0$
      for v $:= 1, \ldots, n$ **do**
        $TEMP := TEMP + |U_{lv}|$
      end for
      if $TEMP > MAX$ **then**
        $MAX := TEMP$
        $ARGMAX := l$
      end if
    end for
    for i $:= 1, \ldots, k$ **do**
      for v $:= 1, \ldots, n$ **do**
        $S_i := S_i + 5n \cdot \operatorname{sgn}(\beta_i^t) \cdot \|(A_i)_{MAX,v}|$// compute subgradient for max row
      end for
    end for
  end if
  for i $:= 1, \ldots, k$ **do**
    $\beta_i^{t+1} = \beta_i^t - \eta S_i$ // Iteration of subgradient descent
  end for
end for
Output $\hat{\beta} := \frac{1}{T}\sum_{t=1}^{T} \beta^t$

subject to the constraint $\|A_{\boldsymbol{\beta}}\|_\infty \leq M$. Thus, in order to solve 10, we can equivalently solve

$$\arg \min_{\boldsymbol{\beta}: \|A_{\boldsymbol{\beta}}\|_\infty \leq M} \psi(\boldsymbol{\beta}) \tag{46}$$

One way to solve this is to use constrained optimization. However, we would like to avoid the complicated projection procedure associated with this strategy. We note that projecting to the set of matrices with low infinity norm does not directly translate to a similar procedure in the space of **fi**, since A_i might have large infinity norm. Instead, we will solve the following regularized optimization problem:

$$\arg \min_{\boldsymbol{\beta}} \left(\psi(\boldsymbol{\beta}) + \lambda \max(0, \|A_{\boldsymbol{\beta}}\|_\infty - M) \right)$$

Set $h(\boldsymbol{\beta}) = \psi(\boldsymbol{\beta}) + \lambda \max(0, \|A_{\boldsymbol{\beta}}\|_\infty - M)$, which is clearly a convex function. We will describe a way to pick λ shortly after. Denote by $\boldsymbol{\beta}_1$ the solution of this optimization problem and by $\boldsymbol{\beta}^*$ the one corresponding to J^*. Also, suppose the algorithm outputs a point $\hat{\boldsymbol{\beta}}$ for accuracy ϵ. The details are given in Algorithm **??**. We will prove that $\|A_{\boldsymbol{\beta}_1}\|_\infty, \|A_{\hat{\boldsymbol{\beta}}}\|_\infty \leq 3M$. First, for $\boldsymbol{\beta} = \mathbf{0}$ we have $\|A_{\mathbf{0}}\|_\infty \leq M$ and by equation 11 we get $\psi(\mathbf{0}) = n \log 2$.

We set $\lambda = 5n$. We now examine what the value of the function is $\boldsymbol{\beta}$ such that $\|A_{\boldsymbol{\beta}}\|_\infty \geq 3M$. If we manage to show that it is greater than the value at 0, then it is clear that the minimizer will not lie in this set. Using the inequality $\cosh(x) \geq \exp -|x|$ and equation 11 we have:

$$\psi(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\log \cosh \left((A_{\boldsymbol{\beta}})_i x \right) - x_i (A_{\boldsymbol{\beta}})_i x + \log 2 \right) \geq \sum_{i=1}^{n} \left(-|(A_{\boldsymbol{\beta}})_i x| - x_i (A_{\boldsymbol{\beta}})_i x + \log 2 \right)$$

Using the triangle inequality and the fact that x is a $\{+1, -1\}^n$ vector, we have: $|(A_{\boldsymbol{\beta}})_i x| \leq \|A_{\boldsymbol{\beta}}\|_\infty$. Also, since $\|A_{\boldsymbol{\beta}}\|_2 \leq \|A_{\boldsymbol{\beta}}\|_\infty$ we have:

$$\sum_{i=1}^{n} x_i (A_{\boldsymbol{\beta}})_i x = x^\top A_{\boldsymbol{\beta}} x \leq \|A_{\boldsymbol{\beta}}\|_2 \|x\|_2^2 = n \|A_{\boldsymbol{\beta}}\|_2 \leq n \|A_{\boldsymbol{\beta}}\|_\infty$$

Overall, we get:

$$\psi(\boldsymbol{\beta}) \geq -n\|A_{\boldsymbol{\beta}}\|_\infty - n\|A_{\boldsymbol{\beta}}\|_\infty + n \log 2 = -2n\|A_{\boldsymbol{\beta}}\|_\infty + n \log 2$$

So, the value of the optimized function at $\boldsymbol{\beta}$ is

$$h(\beta) = \psi(\beta) + 5n(\|A_{\boldsymbol{\beta}}\|_\infty - M) \geq -2n\|A_{\boldsymbol{\beta}}\|_\infty + n \log 2 + 3n\|A_{\boldsymbol{\beta}}\|_\infty = n \log 2 = h(\mathbf{0}) + n\|A_{\boldsymbol{\beta}}\|_\infty$$

Hence, we conclude that $\|A_{\boldsymbol{\beta}_1}\|_\infty \leq 3M$. Moreover, we have

$$h(\hat{\boldsymbol{\beta}}) - h(\boldsymbol{\beta}_1) \leq \epsilon \leq Mn \implies h(\hat{\boldsymbol{\beta}}) \leq Mn + n \log 2 \implies \|A_{\hat{\boldsymbol{\beta}}}\|_\infty \leq 3M$$

As for the guarantee of the algorithm, we notice that:

$$\psi(\hat{\boldsymbol{\beta}}) \leq h(\hat{\boldsymbol{\beta}}) \leq h(\boldsymbol{\beta}_1) + \epsilon \leq h(\boldsymbol{\beta}^*) + \epsilon = \psi(\boldsymbol{\beta}^*) + \epsilon$$

since $\|A^*_{\boldsymbol{\beta}}\|_\infty \leq M$. Thus, if we denote by $\hat{J} = A_{\hat{\boldsymbol{\beta}}}$, we conclude that:

$$\varphi(\hat{J}) \leq \varphi(J^*) + \epsilon$$

which is what we wanted to prove. It remains now to argue about the computational complexity of the procedure. The algorithm we will use is subgradient descent, which performs reasonably well when the subgradient is upper bounded. The following Theorem can be found in [Bub+15].

Lemma 22 ([Bub+15]). *Suppose f is a convex function with minimum x^* and $g \in \partial f$ is a subgradient such that $\|g\| \leq L$. Suppose we are running the following iterative procedure:*

$$x_{t+1} = x_t - \eta g(x_t)$$

with $\|x_1 - x^*\| \leq R$. If we choose $\eta = R/(L\sqrt{t})$ then

$$f\left(\frac{1}{t}\sum_{i=1}^{t} f(x_i)\right) - f(x^*) \leq \frac{RL}{\sqrt{t}}$$

Hence, if we want accuracy ϵ, we need to run the algorithm for $t = O(R^2L^2/\epsilon^2)$ iterations. We now argue that both R and L are polynomially bounded in our case. To bound a subgradient of h, we begin by bounding each partial derivative of ψ, since for this function it is also a subgradient. First of all, it is easy to see that $\partial\psi(\boldsymbol{\beta})/\partial\beta_i \leq 2\sqrt{n}$. Indeed, by equation 12 we get:

$$\begin{aligned}
\left|\frac{\psi(\boldsymbol{\beta})}{\partial\beta_i}\right| &= \left|\sum_{k=1}^{n}((A_i)_k x)\left(\tanh((A_{\boldsymbol{\beta}})_k x) - x_k\right)\right| \leq \sum_{k=1}^{n}|(A_i)_k x|\left|\tanh((A_{\boldsymbol{\beta}})_k x) - x_k\right| \\
&\leq \sum_{k=1}^{n}|(A_i)_k x|\left(|\tanh((A_{\boldsymbol{\beta}})_k x)| + 1\right) \leq 2\sum_{k=1}^{n}|(A_i)_k x| \\
&\leq 2\sum_{k=1}^{n}\sum_{j=1}^{n}|(A_i)_{kj}| \leq 2\sqrt{n}\|A_i\|_F = 2\sqrt{n}
\end{aligned}$$

In the preceding calculation, we used the Cauchy Schwarz inequality along with the fact that A_i belongs to the orthonormal basis. Now, we turn our attention to the non-smooth part of h. We use the following general fact about subgradients.

Lemma 23 (folklore). *Suppose $f_1, f_2 : \mathbb{R}^n \mapsto \mathbb{R}$ are differentiable convex functions. Then, $h = \max(f_1, f_2)$ has a subgradient of the form*

$$g(x) = \left\{ \begin{array}{l} \nabla f_1(x) \text{ , if } f_1(x) \geq f_2(x) \\ \nabla f_2(x) \text{ , otherwise} \end{array} \right\}$$

This means that to bound the subgradient of $\lambda\max(0, \|A_{\boldsymbol{\beta}}\|_\infty - M)$ we just need to bound the gradient of each function in the max. The function 0 obviously has gradient 0. We have

$$\|A_{\boldsymbol{\beta}}\|_\infty = \max_{u\in[n]}\sum_{v=1}^{n}|(A_{\boldsymbol{\beta}})_{uv}|$$

Hence, to bound the subgradient of this function, we focus on a fixed row u. Set

$$L(\boldsymbol{\beta}) = \sum_{v=1}^{n}|(A_{\boldsymbol{\beta}})_{uv}| = \sum_{v=1}^{n}\left|\sum_{i=1}^{k}\beta_i(A_i)_{uv}\right|$$

. Using the fact that a subgradient for $|x|$ is $\text{sgn}(x)$, we obtain that a subgradient of L w.r.t. variable β_i is $\text{sgn}(\beta_i)\sum_{v=1}^{n}|(A_i)_{uv}|$, which is clearly bounded in norm by $\sum_{v=1}^{n}|(A_i)_{uv}| \leq \sqrt{n}$. Hence, the subgradient w.r.t. β_i of the function $\lambda\max(0, \|A_{\boldsymbol{\beta}}\|_\infty - M)$ is bounded in absolute

value by $\lambda\sqrt{n} = 5n\sqrt{n}$. This means that the subgradient g_i of h is $O(n\sqrt{n})$ in absolute value. It follows that $\|g\|_2 \leq n\sqrt{k}\sqrt{n}$. Finally, by choosing $x_1 = 0$ we have

$$\|x_1 - \boldsymbol{\beta}^*\| = \|A_{x_1} - A_{\boldsymbol{\beta}^*}\|_F \leq \sqrt{n}\|A_{x_1} - A_{\boldsymbol{\beta}^*}\|_\infty \leq M\sqrt{n}$$

Hence, after $t = O(n^4M^2k/\epsilon^2)$ rounds we achieve error of at most ϵ. The only part of the algorithm we haven't analysed is the Gram-Schmidt computation. It is well known that for a subspace of $\mathbb{R}^k$ of dimension l, the Gram-Schmidt procedure takes $O(ml^2)$. For $m = k, l = n^2$ we see that this is less than the complexity for the optimization part.

The last issue we should address to get the time complexity of the algorithm is the calculation of the subgradient. The gradient of ψ amounts to the calculation of $A_{\boldsymbol{\beta}}x$, which takes $O(n^2k)$ time(A_ix can be precomputed for all i). The calculation of the subgradient for the nonsmooth part is easy once we determine which row has the maximum absolute sum for the particular value of $\boldsymbol{\beta}$. This also takes $O(kn^2)$ time, since we have to calculate the sum of the matrices in each row. Once we do that, a precomputation of the row sums of each matrix can give us the subgradient in $O(1)$ time. Overall, computation of the gradient takes $O(kn^2)$ time, which means that our algorithm runs in $O(k^2n^6M^2/\epsilon^2)$ time. □

B Dobrushin's Uniqueness Condition

Here we present Dobrushin's uniqueness condition in its full generality. First we define the influence of a node j on a node i.

Definition 8 (Influence in Graphical Models)**.** *Let π be a probability distribution over some set of variables V. Let B_j denote the set of state pairs (X, Y) which differ only in their value at variable j. Then the influence of node j on node i is defined as*

$$I(j,i) = \max_{(X,Y)\in B_j} \left\| \pi_i(.|X^{-i}) - \pi_i(.|Y^{-i}) \right\|_{TV}$$

Now, we are ready to state Dobrushin's condition.

Definition 9 (Dobrushin's Uniqueness Condition)**.** *Consider a distribution π defined on a set of variables V. Let*

$$\alpha = \max_{i\in V} \sum_{j\in V} I(j,i)$$

π is said to satisfy Dobrushin's uniqueness condition if $\alpha < 1$.

Notice that $\|J^*\|_\infty < 1$ implies Dobrushin's condition and a proof can found in [Cha05]. A generalization of this condition was given by [Hay06] who defined the generalized Dobrushin's condition as $\|I\|_2 < 1$, where $I = I(i,j)$ is the influence matrix. This condition, while being weaker, retains most desirable properties of the original Dobrushin's condition.

C Technical Proofs

We begin with the proof of Lemma 5 and then move to the proof of Lemma 8.

C.1 Proof of Lemma 5

In this section, we prove the following lemma:

Lemma 5. *For any symmetric matrix A with zeros on the diagonal, we have*

$$\Pr\left[|\psi_j(x;A)| \geq t \;\middle|\; x_{-I_j}\right] \leq \exp\left(-c\min\left(\frac{t^2}{\|\mathbb{E}\left[Ax \;\middle|\; x_{-I_j}\right]\|_2^2 + \|A\|_F^2}, \frac{t}{\|A\|_2}\right)\right).$$

We will decompose $\psi_j(x;A)$ to parts that depend on x_{I_j} and parts the depend on the parameters that we condition on, x_{-I_j}:

$$\psi_j(x;A) = \sum_{i\in I_j}(A_{i,I_j}x_{I_j} + A_{i,-I_j}x_{-I_j})(x_i - \tanh(J^*_{i,I_j}x_{I_j} + J^*_{i,-I_j}x_{-I_j})) = \sum_{i\in I_j}(A'_ix' + b'_i)(x'_i - \tanh(J'_ix' + h'_i)), \tag{47}$$

where

$$A' = A_{I_jI_j}; \quad b' = A_{I_j,-I_j}x_{-I_j}; \quad J' = J^*_{I_j,I_j}; \quad h' = J^*_{I_j,-I_j}x_{-I_j}; \quad \text{and } x' = x_{I_j}.$$

Notice that A', b', J', h' are fixed conditioned on x_{-I_j} and x' is distributed as the conditional distribution of x_{I_j} conditioned on x_{-I_j}. Furthermore, x' is a $(1/2,\gamma)$-Ising model, with interaction matrix J' and external field h', conditioned on x_{-I_j}. Hence, the following lemma will imply that the right hand size of (47) concentrates conditioned on x_{-I_j}.

Lemma 24. *Let x be a $(1/2,\gamma)$ Ising model over $\{-1,1\}^m$ with interaction matrix J and external field h. Let A be a symmetric real matrix of dimension $m\times m$ with zeros on the diagonal, let $b\in\mathbb{R}^m$ be a vector and let*

$$f(x) = \sum_{i\in[m]}(A_ix + b_i)(x - \tanh(J_ix + h)).$$

Then, for any $t > 0$,

$$\Pr[|f(x)| \geq t] \leq \exp\left(-c\min\left(\frac{t^2}{\|\mathbb{E}Ax+b\|_2^2}, \frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|_2}\right)\right),$$

where $c > 0$ is lower bounded by a constant constant whenever $\|A\|_\infty$, $\|b\|_\infty$, $\|A'\|_\infty$ and $\|b'\|_\infty$ are bounded from above by a constant.

First we derive Lemma 5 based on Lemma 24. Substituting $A = A'$, $b = b'$, $J = J'$ and $h = h'$ and $x = x'$, we derive that,

$$\begin{aligned}
\Pr\left[\psi_j(x;A) \geq t \;\middle|\; x_{-I_j}\right] &\leq \exp\left(-c\min\left(\frac{t^2}{\|\mathbb{E}A'x+b'\|_2^2}, \frac{t^2}{\|A'\|_F^2}, \frac{t}{\|A'\|_2}\right)\right)\\
&= \exp\left(-c\min\left(\frac{t^2}{\|\mathbb{E}A_{I_j}x\|_2^2}, \frac{t^2}{\|A_{I_j,I_j}\|_F^2}, \frac{t}{\|A_{I_j,I_j}\|_2}\right)\right)\\
&\leq \exp\left(-c\min\left(\frac{t^2}{\|\mathbb{E}Ax\|_2^2}, t^2, \frac{t}{\|A\|_2}\right)\right),
\end{aligned}$$

using the fact that $\|A_{I_j,I_j}\|_2 \leq \|A\|_2$ and that $\|A_{I_j,I_j}\|_F \leq \|A\|_F = 1$ for all $A\in\mathcal{A}$. This concludes the proof of Lemma 5. Lastly, we prove Lemma 24.

Proof of Lemma 24. By abuse of notation, define $\tanh\colon \mathbb{R}^m \to \mathbb{R}^m$ by

$$\tanh(y_1, \ldots, y_m) = (\tanh(y_1), \ldots, \tanh(y_m)).$$

Decompose $f(x)$ as

$$f(x) = g(x)^\top h(x),$$

where $g(x) = Ax + b$ and $h(x) = x - \tanh(Jx + h)$.

We start by bounding $\sum_i (D_i f(x))^2$. Decompose

$$\begin{aligned} 2D_i f(x) &= f(x_{i+}) - f(x_{i-}) = g(x_{i+})^\top (h(x_{i+}) - h(x_{i-})) + (g(x_{i+}) - g(x_{i-}))^\top h(x_{i-}) \\ &= (g(x_{i+}) - g(x))^\top (h(x_{i+}) - h(x_{i-})) + g(x)^\top (h(x_{i+}) - h(x_{i-})) \\ &+ (g(x_{i+}) - g(x_{i-}))^\top (h(x_{i-}) - h(x)) + (g(x_{i+}) - g(x_{i-}))^\top h(x). \end{aligned}$$

By Cauchy Schwartz, for any $a, b, c, d \in \mathbb{R}$, we have that $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$. Hence,

$$\begin{aligned} \sum_{i=1}^m (D_i f(x))^2 &\le \sum_{i=1}^m ((g(x_{i+}) - g(x))^\top (h(x_{i+}) - h(x_{i-})))^2 + \sum_{i=1}^m (g(x)^\top (h(x_{i+}) - h(x_{i-})))^2 \\ &+ \sum_{i=1}^m ((g(x_{i+}) - g(x_{i-}))^\top (h(x_{i-}) - h(x)))^2 + \sum_{i=1}^m ((g(x_{i+}) - g(x_{i-}))^\top h(x))^2. \end{aligned} \tag{48}$$

We will bound the terms in the right hand side of (48) one after the other.

Term 4:

$$(g(x_{i+}) - g(x_{i-}))^\top h(x) = (x_{i+} - x_{i-})^\top A(x - \tanh(Jx + h)) = 2e_i^\top A(x - \tanh(Jx + h)).$$

Summing over all i, we get

$$\sum_i ((g(x_{i+}) - g(x_{i-}))^\top h(x))^2 = 4\|A(x - \tanh(Jx + h))\|_2^2.$$

Term 3: using the fact that tanh is 1-Lipschitz and Cauchy Schwartz,

$$\begin{aligned} |(g(x_{i+}) - g(x_{i-}))^\top (h(x_{i-}) - h(x))| &= |2e_i^\top A(x_{i-} - \tanh(Jx_{i-} + h) - x - \tanh(Jx + h))| \\ &\le 2\|e_i^\top A\|_2 \|x_{i-} - \tanh(Jx_{i-} + h) - x - \tanh(Jx + h)\|_2 \\ &\le 2\|e_i^\top A\|_2 (\|x_{i-} - x\|_2 + \|\tanh(Jx_{i-} + h) - \tanh(Jx + h)\|_2) \\ &\le 2\|e_i^\top A\|_2 (2 + \|Jx_{i-} + h - Jx - h\|_2) \\ &\le 2\|e_i^\top A\|_2 (2 + \|J\|_2 \|x_{i-} - x\|_2) \\ &\le C\|e_i^\top A\|_2. \end{aligned}$$

Summing over all i, we get

$$\sum_i |(g(x_{i+}) - g(x_{i-}))^\top (h(x_{i-}) - h(x))|^2 \le C \sum_i \|e_i^\top A\|_2^2 = C\|A\|_F^2.$$

Term 1:

$$|(g(x_{i+}) - g(x))^\top (h(x_{i+}) - h(x_{i-}))| = |(x_{i_+} - x)^\top A(x_{i+} - \tanh(Jx_{i+} + h) - x_{i-} - \tanh(Jx_{i-} + b))|$$
$$\leq \|(x_{i_+} - x)^\top A\|_2 \|x_{i+} - \tanh(Jx_{i+} + h) - x_{i-} - \tanh(Jx_{i-} + b)\|_2 \leq C\|e_i^\top A\|_2,$$

using a bound similar to the above. Summing over all i we get

$$\sum_i |(g(x_{i+}) - g(x))^\top (h(x_{i+}) - h(x_{i-}))|^2 \leq C\|A\|_F^2.$$

Term 2:

$$\sum_i (g(x)^\top (h(x_{i+}) - h(x_{i-})))^2 = \|Wg(x)\|_2^2,$$

where W is a matrix of size $m \times m$ such that

$$W_{ij} = h(x_{i+})_j - h(x_{i-})_j = (x_{i+})_j - (x_{i-})_j - \tanh(J_j^\top x_{i+} + h) + \tanh(J_j^\top x_{i-} + h). \tag{49}$$

Using the Lipschitzness of tanh and the triangle inequality,

$$|W_{ij}| \leq |(x_{i+})_j - (x_{i-})_j| + |J_j^\top (x_{i+} - x_{i-})| = 2(\mathbf{1}(i = j) + |J_{ij}|).$$

We obtain that

$$\|W\|_2 \leq \|W\|_\infty \leq 2\|J\|_\infty + 2 \leq C. \tag{50}$$

Hence, $\|Wg(x)\|_2^2 \leq \|W\|_2^2 \|g(x)\|_2^2 \leq C^2\|Ax + b\|_2^2$.

To summarize, by (48) and the calculations below, we obtain that

$$\sum_i (D_i f(x))^2 \leq C(\|A\|_F^2 + \|Ax + b\|_2^2 + \|A(x - \tanh(Jx + h)))\|_2^2.$$

Define the pseudo discrete derivative to be a function of $2m + 1$ coordinates, such that for coordinate i, $i \in [m]$, we have $\widetilde{D}_i(x) = C(A_i^\top x + b_i)$, in coordinate $m + i$ we have $\widetilde{D}_{n+i}(x) = CA_i^\top (x - \tanh(Jx + h))$, and in coordinate $2m + 1$ we have $\widetilde{D}_{2m+1}(x) = C\|A\|_F$.

Next, we like to define a pseudo discrete Hessian. For this purpose, we bound $\sum_{j=1}^m D_j(\xi^\top \widetilde{D}(x))^2$, for any fixed $\xi \in \mathbb{R}^{2m+1}$. Note that using the fact that $\widetilde{D}_{2m+1}$ is constant in x and using Cauchy Schwartz,

$$\sum_{j=1}^m D_j(\xi^\top \widetilde{D}(x))^2 = \sum_{j=1}^m \left(\sum_{i=1}^{2m+1} \xi_i (\widetilde{D}_i(x_{j+}) - \widetilde{D}_i(x_{j-})) \right)^2$$
$$\leq 2\sum_{j=1}^m \left(\sum_{i=1}^{m} \xi_i (\widetilde{D}_i(x_{j+}) - \widetilde{D}_i(x_{j-})) \right)^2 + 2\sum_{j=1}^m \left(\sum_{i=n+1}^{2m} \xi_i (\widetilde{D}_i(x_{j+}) - \widetilde{D}_i(x_{j-})) \right)^2. \tag{51}$$

We will bound both terms from the right hand size of (51). Starting with the first term, for any $i \in [m]$ we have

$$\widetilde{D}_i(x_{j+}) - \widetilde{D}_i(x_{j-}) = 2A_{ij}.$$

Hence,

$$\sum_{j=1}^{m}\left(\sum_{i=1}^{m}\xi_i(\widetilde{D}_i(x_{j+})-\widetilde{D}_i(x_{j-}))\right)^2=4\sum_j\left(\sum_i\xi_iA_{ij}\right)^2=4\|\xi_{1\cdots m}^\top A\|_2^2,$$

where $\xi_{1\cdots m}=(\xi_1,\ldots,\xi_m)$. For the second term of (51), we have:

$$\sum_{j=1}^{m}\left(\sum_{i=n+1}^{2m}\xi_i(\widetilde{D}_i(x_{j+})-\widetilde{D}_i(x_{j-}))\right)^2$$
$$=\left\|\xi_{m+1\cdots 2m}^\top A\left(\sum_{i\in[n]}x_{i+}-x_{i-}-\tanh(A'x_{i+}+h)+\tanh(A'x_{i-}+h)\right)\right\|_2^2=\|\xi_{m+1\cdots 2m}^\top AW\|_2^2,$$

for the matrix W that we defined in the calculations of the discrete derivative, in (49). By (50) we get that

$$\|\xi_{m+1\cdots 2m}^\top AW\|_2^2\le\|\xi_{m+1\cdots 2m}^\top A\|_2^2\|W\|_2^2\le C\|\xi_{m+1\cdots 2m}^\top A\|_2^2.$$

By (51) and the bounds on both terms in its right hand side, we derive that

$$\sum_{j\in[m]}D_j(\xi^\top\widetilde{D}(x))^2\le\|\xi^\top\widetilde{H}\|_2^2,$$

where $\widetilde{H}=C(A|A|0)^\top$ is the matrix of dimension $(2m+1)\times m$ obtained from stacking two copies of A one on top of each other on top of one row of zeros at the bottom, all multiplied by a sufficiently large constant C. Hence, we can define the pseudo Hessian as the constant function $\widetilde{H}(x)=\widetilde{H}$.

Lastly, we would like to apply Theorem 6, applying it with the pseudo discrete derivative and Hessian defined above. We would just have to calculate:

$$\|\mathbb{E}_x[\widetilde{D}(x)]\|_2^2=C\sum_{i=1}^{m}(\mathbb{E}_x[A_i^\top x+b_i])^2+C\sum_{i=1}^{m}(\mathbb{E}_x[A_i^\top(x-\tanh(Jx+h))])^2+C\|A\|_F^2.$$

The first summand equals $\|\mathbb{E}[Ax+b]\|_2^2$, while the second equals zero, from the same argument as in Claim 1. This implies that $\|\mathbb{E}[\widetilde{D}(x)]\|_2^2\le C\|\mathbb{E}[Ax+b]\|_2^2+C\|A\|_F^2$. Next, we bound the terms corresponding to the pseudo Hessian: we have that $\|\widetilde{H}\|_2\le C\|A\|_2$, and $\|\widetilde{H}\|_F^2\le C\|A\|_F^2$. Plugging these in Theorem 6 concludes the proof. □

C.2 Proof of Lemma 8

Now, we move on to the proof of Lemma 8. The function that we wish to show concentration about is a second degree polynomial, hence Theorem 5 applies. However, this Theorem requires the matrix to have 0 in the diagonal, which is not necessarily the case for $A^\top A$. Hence, we need to modify the matrix so that it is zero-diagonal, obtain the concentration bound for the modified matrix and then translate the result in terms of the original matrix. This is done in the following proof.

Proof of Lemma 8. Denote $p(x)=\|Ax\|_2^2=x^\top A^\top Ax$. Let E be obtained from $A^\top A$ by zeroing all elements of the diagonal and denote $\widetilde{p}(x)=x^\top Ex$. Note that $\widetilde{p}(x)-p(x)$ is a constant as a

function of x, since $x_i^2 = 1$ for all i. This means that it suffices to bound the deviation of $\widetilde{p}(x)$. By Lemma 2, conditioning on x_{-I_j} yields an Ising model with Dobrushin constant $1/2$. Hence, we can apply Theorem 5 and get that for any $t \geq 0$,

$$\Pr\left[\left|\widetilde{p}(x) - \mathbb{E}\left[\widetilde{p}(x) \;\middle|\; x_{-I_j}\right]\right| > t \;\middle|\; x_{-I_j}\right] \leq \exp\left(-c\min\left(\frac{t^2}{\|E\|_F^2 + \|\mathbb{E}[Ex|x_{-I_j}]\|_2^2}, \frac{t}{\|E\|_2}\right)\right). \quad (52)$$

Now, we show how E can be replaced with $A^\top A$ in (52). First:

$$\left\|\mathbb{E}\left[Ex \;\middle|\; x_{-I_j}\right]\right\|_2 \leq \left\|\mathbb{E}\left[A^\top Ax \;\middle|\; x_{-I_j}\right]\right\|_2 + \left\|\mathbb{E}\left[(E - A^\top A)x \;\middle|\; x_{-I_j}\right]\right\|_2,$$

which, using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, implies that

$$\begin{aligned}\|E\|_F^2 + \left\|\mathbb{E}\left[Ex \;\middle|\; x_{-I_j}\right]\right\|_2^2 &\leq 2\|E\|_F^2 + 2\left\|\mathbb{E}\left[A^\top Ax \;\middle|\; x_{-I_j}\right]\right\|_2 + 2\left\|\mathbb{E}\left[(E - A^\top A)x \;\middle|\; x_{-I_j}\right]\right\|_2^2 \\ &= 2\left\|A^\top A\right\|_F^2 + 2\left\|\mathbb{E}\left[A^\top Ax \;\middle|\; x_{-I_j}\right]\right\|_2^2.\end{aligned}$$

In the last equality, we used the fact that $\left\|\mathbb{E}\left[(E - A^\top A)x \;\middle|\; x_{-I_j}\right]\right\|_2^2$ is just the sum of the squares of the diagonal entries of $A^\top A$, which means that together with $\|E\|_F^2$ they add up to $\|A\|_F^2$. Next, notice that $\|E\|_2 \leq \|A^\top A\|_2$. To prove this, we note that for all $x \in \mathbb{R}^n$,

$$x^\top Ex = x^\top A^\top Ax - \sum_{i\in[n]} (A^\top A)_{ii} \leq x^\top A^\top Ax.$$

Putting all of this together, we obtain that the right hand side of 52 is bounded by

$$\exp\left(-c'\min\left(\frac{t^2}{\|A^\top A\|_F^2 + \left\|\mathbb{E}\left[A^\top Ax \;\middle|\; x_{-I_j}\right]\right\|_2^2}, \frac{t}{\|A^\top A\|_2}\right)\right).$$

Finally, we want to make this bound depend on A rather than $A^\top A$. First,

$$\|A^\top A\|_F^2 \leq \|A\|_F^2 \|A\|_2^2$$

using the well known inequality $\|AB\|_F \leq \|A\|_2 \|B\|_F$.

Next, we have:

$$\left\|\mathbb{E}\left[A^\top Ax \;\middle|\; x_{-I_j}\right]\right\|_2^2 = \left\|A^\top \mathbb{E}\left[Ax \;\middle|\; x_{-I_j}\right]\right\|_2^2 \leq \|A\|_2^2 \left\|\mathbb{E}\left[Ax \;\middle|\; x_{-I_j}\right]\right\|_2^2 \quad (53)$$

Lastly,

$$\|A^\top A\|_2 = \|A\|_2^2.$$

This concludes the proof. □

References

[Ada+19] R. Adamczak, M. Kotowski, B. Polaczyk, M. Strzelecki, et al. "A note on concentration for polynomials in the Ising model". In: *Electronic Journal of Probability* 24 (2019).

[AF12] L. Anselin and R. Florax. *New directions in spatial econometrics*. Springer Science & Business Media, 2012.

[Ans01] L. Anselin. "Spatial econometrics". In: *A companion to theoretical econometrics* 310330 (2001).

[Ans13] L. Anselin. *Spatial econometrics: methods and models*. Vol. 4. Springer Science & Business Media, 2013.

[AS04] N. Alon and J. H. Spencer. *The probabilistic method*. John Wiley & Sons, 2004.

[BDF09] Y. Bramoullé, H. Djebbari, and B. Fortin. "Identification of peer effects through social networks". In: *Journal of econometrics* 150.1 (2009), pp. 41–55.

[Ber+09] P. Berti, I. Crimaldi, L. Pratelli, P. Rigo, et al. "Rate of convergence of predictive distributions for dependent data". In: *Bernoulli* 15.4 (2009), pp. 1351–1367.

[Bes74] J. Besag. "Spatial interaction and the statistical analysis of lattice systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 192–225.

[Bes75] J. Besag. "Statistical analysis of non-lattice data". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 24.3 (1975), pp. 179–195.

[BLM00] M. Bertrand, E. F. Luttmer, and S. Mullainathan. "Network effects and welfare cultures". In: *The Quarterly Journal of Economics* 115.3 (2000), pp. 1019–1055.

[BM18] B. B. Bhattacharya and S. Mukherjee. "Inference in Ising models". In: *Bernoulli* 24.1 (2018), pp. 493–525.

[BN18] G. Bresler and D. Nagaraj. "Optimal single sample tests for structured versus unstructured network data". In: *arXiv preprint arXiv:1802.06186* (2018).

[BN+19] G. Bresler, D. Nagaraj, et al. "Stein's method for stationary distributions of Markov chains and application to Ising models". In: *The Annals of Applied Probability* 29.5 (2019), pp. 3230–3265.

[Bre15] G. Bresler. "Efficiently learning Ising models on arbitrary graphs". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015, pp. 771–782.

[Bub+15] S. Bubeck et al. "Convex optimization: Algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[CF13] N. A. Christakis and J. H. Fowler. "Social contagion theory: examining dynamic social networks and human behavior". In: *Statistics in medicine* 32.4 (2013), pp. 556–577.

[Cha05] S. Chatterjee. "Concentration inequalities with exchangeable pairs (Ph. D. thesis)". In: *arXiv preprint math/0507526* (2005).

[Cha07] S. Chatterjee. "Estimation in spin glasses: A first step". In: *The Annals of Statistics* 35.5 (2007), pp. 1931–1946.

[CL68] C Chow and C. Liu. "Approximating discrete probability distributions with dependence trees". In: *IEEE transactions on Information Theory* 14.3 (1968), pp. 462–467.

[CVV19] J. Y. Chen, G. Valiant, and P. Valiant. "How bad is worst-case data if you know where it comes from?" In: *arXiv* abs/1911.03605 (2019).

[Dag+19] Y. Dagan, C. Daskalakis, N. Dikkala, and S. Jayanti. "Learning from Weakly Dependent Data under Dobrushin's Condition". In: *Conference on Learning Theory*. 2019, pp. 914–928.

[DDK17] C. Daskalakis, N. Dikkala, and G. Kamath. "Concentration of multilinear functions of the Ising model with applications to network data". In: *Advances in Neural Information Processing Systems*. 2017, pp. 12–23.

[DDP19] C. Daskalakis, N. Dikkala, and I. Panageas. "Regression from dependent observations". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 881–889.

[DMR11] C. Daskalakis, E. Mossel, and S. Roch. "Evolutionary Trees and the Ising Model on the Bethe Lattice: A Proof of Steel's Conjecture". In: *Probability Theory and Related Fields* 149.1 (2011), pp. 149–189.

[DS03] E. Duflo and E. Saez. "The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment". In: *The Quarterly journal of economics* 118.3 (2003), pp. 815–842.

[DS87] R. Dobrushin and S. Shlosman. "Completely analytical interactions: constructive description". In: *Journal of Statistical Physics* 46.5-6 (1987), pp. 983–1014.

[Duc16] J. Duchi. "Lecture notes for statistics 311/electrical engineering 377". In: *URL: https://stanford. edu/class/stats311/Lectures/full_notes. pdf. Last visited on* 2 (2016), p. 23.

[Ell93] G. Ellison. "Learning, Local Interaction, and Coordination". In: *Econometrica* 61.5 (1993), pp. 1047–1071.

[Fel04] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates Sunderland, 2004.

[GG86] S. Geman and C. Graffigne. "Markov Random Field Image Models and their Applications to Computer Vision". In: *Proceedings of the International Congress of Mathematicians*. American Mathematical Society, 1986, pp. 1496–1517.

[GLP17] R. Gheissari, E. Lubetzky, and Y. Peres. "Concentration inequalities for polynomials of contracting Ising models". In: *arXiv preprint arXiv:1706.00121* (2017).

[GLP+18] R. Gheissari, E. Lubetzky, Y. Peres, et al. "Concentration inequalities for polynomials of contracting Ising models". In: *Electronic Communications in Probability* 23 (2018).

[GM18] P. Ghosal and S. Mukherjee. "Joint estimation of parameters in Ising model". In: *arXiv preprint arXiv:1801.06570* (2018).

[GSS19] F. Götze, H. Sambale, and A. Sinulis. "Higher order concentration for functions of weakly dependent random variables". In: *Electronic Journal of Probability* 24 (2019).

[GSS96] E. L. Glaeser, B. Sacerdote, and J. A. Scheinkman. "Crime and social interactions". In: *The Quarterly Journal of Economics* 111.2 (1996), pp. 507–548.

[Hay06] T. P. Hayes. "A simple condition implying rapid mixing of single-site dynamics on spin systems". In: *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. IEEE. 2006, pp. 39–46.

[HKM17] L. Hamilton, F. Koehler, and A. Moitra. "Information theoretic properties of Markov random fields, and their algorithmic applications". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2463–2472.

[Isi25] E. Ising. "Beitrag zur theorie des ferromagnetismus". In: *Zeitschrift für Physik* 31.1 (1925), pp. 253–258.

[KM15] V. Kuznetsov and M. Mohri. "Learning theory and algorithms for forecasting non-stationary time series". In: *Advances in neural information processing systems*. 2015, pp. 541–549.

[KM17] A. Klivans and R. Meka. "Learning graphical models using multiplicative weights". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 343–354.

[Km79] R. Z. Khas' minskii. "A lower bound on the risks of non-parametric estimates of densities in the uniform metric". In: *Theory of Probability & Its Applications* 23.4 (1979), pp. 794–798.

[KR+08] L. A. Kontorovich, K. Ramanan, et al. "Concentration inequalities for dependent random variables via the martingale method". In: *The Annals of Probability* 36.6 (2008), pp. 2126–2158.

[KR17] A. Kontorovich and M. Raginsky. "Concentration of measure without independence: a unified approach via the martingale method". In: *Convexity and Concentration*. Springer, 2017, pp. 183–210.

[Lat+06] R. Latała et al. "Estimates of moments and tails of Gaussian chaoses". In: *The Annals of Probability* 34.6 (2006), pp. 2315–2331.

[Lau96] S. L. Lauritzen. *Graphical models*. Vol. 17. Clarendon Press, 1996.

[LeS08] J. P. LeSage. "An introduction to spatial econometrics". In: *Revue d'économie industrielle* 123 (2008), pp. 19–44.

[LHG16] B. London, B. Huang, and L. Getoor. "Stability and generalization in structured prediction". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 7808–7859.

[Lon+13] B. London, B. Huang, B. Taskar, and L. Getoor. "Collective stability in structured prediction: Generalization from one example". In: *International Conference on Machine Learning*. 2013, pp. 828–836.

[Man93] C. F. Manski. "Identification of endogenous social effects: The reflection problem". In: *The review of economic studies* 60.3 (1993), pp. 531–542.

[Mar03] K. Marton. "Measure concentration and strong mixing". In: *Studia Scientiarum Mathematicarum Hungarica* 40.1-2 (2003), pp. 95–113.

[MR09] M. Mohri and A. Rostamizadeh. "Rademacher complexity bounds for non-iid processes". In: *Advances in Neural Information Processing Systems*. 2009, pp. 1097–1104.

[MR10] M. Mohri and A. Rostamizadeh. "Stability bounds for stationary φ-mixing and β-mixing processes". In: *Journal of Machine Learning Research* 11.Feb (2010), pp. 789–814.

[MS17] D. J. McDonald and C. R. Shalizi. "Rademacher complexity of stationary sequences". In: *arXiv preprint arXiv:1106.0730* (2017).

[Pes10] V. Pestov. "Predictive PAC learnability: A paradigm for learning from exchangeable input data". In: *Granular Computing (GrC), 2010 IEEE International Conference on*. IEEE. 2010, pp. 387–391.

[RV13] M. Rudelson and R. Vershynin. "Hanson-Wright inequality and sub-gaussian concentration". In: *Electronic Communications in Probability* 18 (2013).

[RWL10] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. "High-dimensional Ising model selection using l1-regularized logistic regression". In: *The Annals of Statistics* 38.3 (2010), pp. 1287–1319.

[Sac01] B. Sacerdote. "Peer effects with random assignment: Results for Dartmouth roommates". In: *The Quarterly journal of economics* 116.2 (2001), pp. 681–704.

[SS14] A. Sly and N. Sun. "Counting in two-spin models on d-regular graphs". In: *The Annals of Probability* 42.6 (2014), pp. 2383–2416.

[SW12] N. P. Santhanam and M. J. Wainwright. "Information-theoretic limits of selecting binary graphical models in high dimensions". In: *IEEE Transactions on Information Theory* 58.7 (2012), pp. 4117–4134.

[SZ92] D. W. Stroock and B. Zegarlinski. "The logarithmic Sobolev inequality for discrete spin systems on a lattice". In: *Communications in Mathematical Physics* 149.1 (1992), pp. 175–193.

[TNP08] J. G. Trogdon, J. Nonnemaker, and J. Pais. "Peer effects in adolescent overweight". In: *Journal of health economics* 27.5 (2008), pp. 1388–1399.

[Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[Vuf+16] M. Vuffray, S. Misra, A. Lokhov, and M. Chertkov. "Interaction screening: Efficient and sample-optimal learning of Ising models". In: *Advances in Neural Information Processing Systems*. 2016, pp. 2595–2603.

[Wei05] D. Weitz. "Combinatorial criteria for uniqueness of Gibbs measures". In: *Random Structures & Algorithms* 27.4 (2005), pp. 445–475.

[WJ+08] M. J. Wainwright, M. I. Jordan, et al. "Graphical models, exponential families, and variational inference". In: *Foundations and Trends® in Machine Learning* 1.1–2 (2008), pp. 1–305.

[WSD19] S. Wu, S. Sanghavi, and A. G. Dimakis. "Sparse logistic regression learns all discrete pairwise graphical models". In: *Advances in Neural Information Processing Systems*. 2019, pp. 8069–8079.

[Yu94] B. Yu. "Rates of convergence for empirical processes of stationary mixing sequences". In: *The Annals of Probability* (1994), pp. 94–116.