Estimating Ising Models in Total Variation Distance

Constantinos Daskalakis COSTIS@CSAIL.MIT.EDU
CSAIL, MIT

Vardis Kandiros AK5484@COLUMBIA.EDU
Data Science Institute, Columbia University

Rui Yao RAYYAO@MIT.EDU
CSAIL, MIT

Editors: Steve Hanneke and Tor Lattimore

Abstract

We consider the problem of estimating Ising models over n variables in Total Variation (TV) distance, given l independent samples from the model. While the statistical complexity of the problem is well-understood Devroye et al. (2020), identifying computationally and statistically efficient algorithms has been challenging. In particular, remarkable progress has occurred in several settings, such as when the underlying graph is a tree Daskalakis and Pan (2021); Bhattacharyya et al. (2021), when the entries of the interaction matrix follow a Gaussian distribution Gaitonde and Mossel (2024); Chandrasekaran and Klivans (2024), or when the bulk of its eigenvalues lie in a small interval Anari et al. (2024a); Koehler et al. (2024), but no unified framework for polynomial-time estimation in TV exists so far.

Our main contribution is a unified analysis of the Maximum Pseudo-Likelihood Estimator (MPLE) for two general classes of Ising models. The first class includes models whose interaction matrix has a bounded operator norm. In particular, we focus on the subclass of models that satisfy the Modified Log-Sobolev Inequality (MLSI), a functional inequality that was introduced to study the convergence of the associated Glauber dynamics to stationarity. In the second class of models, the interaction matrix has bounded infinity norm (or *bounded width*), which is the most common assumption in the literature for structure learning of Ising models. We show how our general results for these classes yield polynomial-time algorithms and optimal or near-optimal sample complexity guarantees in a variety of settings. Our proofs employ a variety of tools from tensorization inequalities to measure decompositions and concentration bounds.

Keywords: Total Variation Distance, Graphical models, Ising model

1. Introduction

Undirected graphical models are a widely used framework for capturing conditional independence structure in a high-dimensional distribution. One of the earliest and most prominent instances of these models are *Ising models*, a family of distributions over n binary variables specified by a symmetric *interaction matrix* $J^* \in \mathbb{R}^{n\times n}$, with zero diagonal, and a vector of *external fields* $h \in \mathbb{R}^n$. In terms of these parameters, a probability distribution is defined over $\{-1,1\}^n$, assigning to each vector $x \in \{-1,1\}^n$ probability

$$\underset{J^*,h}{\mathbf{Pr}}[x] = \frac{1}{Z_{J^*,h}} \cdot \exp\left(\frac{1}{2}x^\top J^* x + h^\top x\right) \quad , \tag{1}$$

where the normalizing constant

$$Z_{J^*,h} := \sum_{x \in \{-1,1\}^n} \exp\left(\frac{1}{2} x^\top J^* x + h^\top x\right)$$

is called the *partition function* of the model. We will focus on the case $h = 0$, so we drop the dependence on h whenever that happens. The matrix J^* can also be thought of as the weighted adjacency matrix of a graph with n nodes. This gives rise to the interpretation of an Ising model as a *Markov Random Field (MRF)* Koller and Friedman (2009); Wainwright et al. (2008), where conditional independence relations between variables are encoded as connectivity properties between nodes in the graph defined by J^*. Since its introduction in statistical physics Lenz (1920), the Ising model has been studied intensely in probability theory, and has found profound applications in a variety of fields, including computer vision, economics, computational biology, and the social sciences; see e.g. Geman and Graffigne (1986); Ellison (1993); Felsenstein (2004); Daskalakis et al. (2006); Montanari and Saberi (2010). Motivated in part by these applications, a common challenge is estimating an Ising model given a number of observations that are assumed to be distributed according to some Ising model, e.g. capturing opinions of individuals in a social network, or traits of species in some phylogenetic tree.

The problem of learning the structure of the underlying graph, i.e. the non-zero entries of J^*, given access to l independent samples from an Ising model has received significant attention, due to its importance in capturing conditional independence properties. Under the assumption that the degree of the graph is bounded by d and the non-zero entries of J^* are both upper and lower bounded in absolute value, a breakthrough result by Bresler Bresler (2015) provided a polynomial-time algorithm for identifying the graph topology, albeit with doubly exponential dependence of the sample-size requirements in the degree d. A flurry of subsequent works Hamilton et al. (2017); Vuffray et al. (2016); Lokhov et al. (2018); Klivans and Meka (2017); Wu et al. (2019); Zhang et al. (2020) have culminated in polynomial-time algorithms that estimate every entry of J^* up to error ϵ, given $l = \Theta(e^{\Theta(d)} \log n/\epsilon^4)$ samples from the model, under the assumption that each row of J^* has infinity norm that is upper-bounded. This matches the information-theoretic lower bound of Santhanam and Wainwright (2012). Thus, the problem of estimating the graph structure of an Ising model is by now well understood under relatively general assumptions.

However, often the purpose of estimation is to make predictions of events for various downstream uses of the model Bresler and Karzand (2020); Daskalakis and Pan (2021); Bhattacharyya et al. (2021). It is clear that the right metric to capture this property is not graph similarity, but the *total variation (TV) distance* (formal definition in Section 2) between the true and the estimated distribution. Information-theoretically, Devroye et al. Devroye et al. (2020) show that $\tilde{O}(n^2/\epsilon^2)$ samples from some Ising model are both necessary and sufficient for estimating an Ising model that is ϵ-close in TV. The algorithm proposed in Devroye et al. (2020) involves exhaustive search over all models and is thus computationally infeasible. This motivates the following natural question.

> *Is there a* polynomial-time *algorithm that uses independent samples from an Ising model and outputs some Ising model that is close to the one providing samples in TV distance?*

While for the problem of structure recovery, there has been significant progress towards computationally efficient and statistically optimal algorithms under fairly general settings, attaining similar

guarantees for the TV estimation problem has been challenging. Remarkable progress has occurred across several different directions, such as when the graph is a tree Daskalakis and Pan (2021); Bhattacharyya et al. (2021), when the interaction matrix is sampled from a Gaussian ensemble Gaitonde and Mossel (2024); Chandrasekaran and Klivans (2024), or when most of the spectrum of the matrix lies in a small interval Anari et al. (2024a); Koehler et al. (2024), but no unified framework exists for statistically and computationally efficient procedures for this task. The main contribution of this work is to provide a refined understanding of a natural, polynomial-time algorithm for Ising model estimation, under broad conditions, and derive from our understanding near-optimal (and in some cases optimal) sample-complexity guarantees in a variety of important, specific settings. In particular, we focus on the so-called *Maximum Pseudo-Likelihood Estimator (MPLE)* (formally defined in Section 4), which was introduced in Besag (1974) and is often employed for statistical estimation of autoregressive models. We study the performance of this polynomial-time computable estimator for two general classes of Ising models, which we review in the next couple of paragraphs.

(i) Bounded Operator Norm Condition: A variety of interesting models from statistical physics, such as the SK model Panchenko (2012), involve interaction matrices whose operator norm is upper bounded by a constant. Since these models exhibit complex behavior and undergo phase transitions, we need to incorporate more information about the distribution. One common way of doing that is by studying the properties of an associated Markov Chain, called *Glauber dynamics*, that has this model as its stationary distribution. In particular, we focus on Ising models that satisfy the so-called *Modified Log-Sobolev Inequality (MLSI)*, which guarantees fast convergence of the Glauber dynamics to the stationary distributionBobkov and Tetali (2006)(for formal definitions, see Section A). This condition has been shown to hold for a variety of Ising models under different assumptions Anari et al. (2021, 2024b); Caputo et al. (2015); Chen and Eldan (2022); Blanca et al. (2022); Chen et al. (2021). It is also known to imply many structural properties for the Ising model distribution, such as fast mixing Caputo (2023) and concentration of measure Marton (2015); Sambale and Sinulis (2019); Götze et al. (2021). We prove a general result about the performance of MPLE for estimating Ising Models with bounded operator norm that satisfy MLSI, which implies optimal or near-optimal sample complexity for learning a variety of Ising models, improving results from prior work.

(ii) Bounded Width Condition: The second class of models we study are *bounded-width* models, where $\|J^*\|_\infty = O(1)$. This has been the canonical class of models considered in most studies of the structure learning problem Santhanam and Wainwright (2012); Bresler (2015); Hamilton et al. (2017); Vuffray et al. (2016); Klivans and Meka (2017); Wu et al. (2019). For this class of models, Klivans and Meka (2017) give a polynomial-time algorithm that uses $O(n^8/\epsilon^4)$ samples and learns an ϵ-multiplicative approximation of $\mathbf{Pr}_{J^*}$, which implies ϵ-closeness in TV. We provide a refined analysis of MPLE, which involves a convex objective that can be optimized efficiently. As a consequence, we show how one could obtain sample complexity guarantees within a single $O(n)$ factor from optimal, assuming the model satisfies a suitable regularity condition that we appropriately define.

Our improved general bounds can be applied in a variety of models, yielding comparable or superior sample complexity to that of prior work. For simplicity, we only discuss examples of models where we get improved complexity guarantees. Table 1 contains a detailed comparison.

At a technical level, our improvements are made possible by using a connection to the problem of *single-sample* estimation of Ising models, which was formulated in Dagan et al. (2021) and implicitly studied elsewhere before; for some references see e.g. Chatterjee (2007); Bhattacharya and

Mukherjee (2018); Daskalakis et al. (2019); Dagan et al. (2021); Mukherjee et al. (2021); Kandiros et al. (2021). This line of work assumes that we are only given a *single* sample from an Ising model whose interaction matrix lies in some low-dimensional subspace, and our goal is to estimate this matrix. These works manage to extract a useful signal from a single sample from the model, even in the presence of strong dependencies within the sample. In contrast, algorithms that rely on multiple samples make strong use of the independence across different samples to employ generalization bounds from learning theory within the estimation procedure. In Dagan et al. (2021), the authors show how to leverage their single-sample estimation methods to obtain a polynomial-time algorithm for learning the structure of an Ising model in the multiple-sample regime. However, their sample complexity is far from optimal. In this paper, we improve upon this single-sample-based approach to obtain various state-of-the-art results in the multiple-sample regime. It is crucial for our development to establish tight upper and lower bounds for the first and second derivatives of the pseudolikelihood function, which we do by using a variety of tools from high-dimensional probability and statistics, such as tensorization inequalities, measure decompositions, and concentration of measure. Along the way, we provide refined guarantees for pseudolikelihood estimation, which could be of independent interest. We believe the connection between single-sample and multiple-sample learning could be more broadly applicable to a variety of estimation problems for Markov Random Fields. Thus, our work serves as a first step towards obtaining efficient and sample-optimal algorithms for TV learning of high-dimensional distributions with complex dependencies.

2. Results

Let $\mathcal{S}_0^n$ denote the set of all symmetric matrices $A \in \mathbb{R}^{n\times n}$ with zeroes on the diagonal. We will make use of the *infinity norm* of a matrix $A \in \mathcal{S}_0^n$, which is defined as $\|A\|_\infty := \max_{i\in[n]} \sum_j |A_{ij}|$, as well as the *operator norm* of A, defined as $\|A\|_2 := \sup_{x\neq 0} \|Ax\|_2/\|x\|_2$. We will occasionally denote this operator norm by $\|A\|_{op}$. Also, the *Frobenius norm* is defined as $\|A\|_F := \sum_i \sum_j A_{ij}^2$. For $A \in \mathcal{S}_0^n$, we denote by A_i the i-th row of A. For a vector $x \in \mathbb{R}^n$ and $i \in [n]$, we denote by x_{-i} the sub-vector obtained by removing the value of coordinate i from the vector. In general, we will use lowercase letters for deterministic quantities and uppercase letters for random quantities. When we refer to an Ising model with interaction matrix J and the external field $\boldsymbol{h}$ is the zero vector, we will write $\mathbf{Pr}_J$ for convenience and simply omit the external field. If we sample $X \sim \mathbf{Pr}_J$, we refer to $X_i \in \{-1, 1\}$ as the *spin* of node i. For two probability measures $\mathbb{P}, \mathbb{Q}$ supported on $\{-1, 1\}^n$, their *Total Variation (TV) Distance* is defined to be $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) := \sup_{A\subseteq\{-1,1\}^n} |\mathbb{P}(A) - \mathbb{Q}(A)|$, where A ranges over all subsets of $\{-1, 1\}^n$.

In this work, we are given l independent samples from an Ising model $\mathbf{Pr}_{J^*}$ as in (1). Our goal will be to properly learn the model in distribution, i.e., to estimate some matrix $\hat{J} \in \mathcal{S}_0^n$ so that $\mathbf{Pr}_{J^*}$ and $\mathbf{Pr}_{\hat{J}}$ are close in TV. Additionally, we would like an algorithm that runs in polynomial time. We now state the results of this investigation for the two different classes of Ising models that we consider.

2.1. Estimating Ising Models with Bounded Operator Norm

A considerable amount of work has focused on identifying classes of Ising models where sampling and inference from the model are computationally tractable. When interactions are strong, it is known that in general these tasks become intractable Sly (2010); Sly and Sun (2012); Galanis et al. (2015). Therefore, a natural direction would be to place limits on the interactions between nodes.

General Class of Models	Applications	Our Work	Prior Work
Bounded 2-norm	Spectrally-bounded models	$\tilde{O}(n^2/\epsilon^2)$ (Corollary 2, optimal)	$O(n^3/\epsilon^4)$ [AJK+24a] [Lee23]
	SK/diluted SK model ($\beta < 0.295...$)	$\tilde{O}(n^4/\epsilon^2)$ (Corollaries 3 and 4)	$\tilde{O}(n^9/\epsilon^8)$ [CK24]
	Antiferromagnetic expanders	$\tilde{O}(n^2/\epsilon^2)$ (Corollary 5, optimal)	$\tilde{O}(n^5/\epsilon^4)$ [Koe+24]
Bounded-width	$(1/\sqrt{n}, 1)$-regular models	$\tilde{O}(n^3/\epsilon^2)$ (Corollary 10)	$\tilde{O}(n^8/\epsilon^4)$ [KM17]
	General	$\tilde{O}(n^4/\epsilon^2)$ (Corollary 8)	$\widetilde{O}(n^8/\epsilon^4)^\dagger$ [KM17]

Table 1: This table contains the sample complexity bound implied by our work, as well as the best known bound from prior work, for the problem of estimating an Ising model to within $\epsilon > 0$ in TV. The best-known prior bound is discussed, where these results are stated.

One of the weakest such constraints would be to assume that the operator norm of the interaction matrix of the model is bounded by some constant. This is satisfied by virtually all examples of Ising models that have been studied in the literature, including the SK model Panchenko (2012) and its diluted variants Talagrand (2010), as well as the Ising model on regular graphs Sly and Sun (2012).

These models often undergo complex phase transition, necessitating additional structural constraints. One natural restriction is to assume that the Glauber dynamics associated with the model converge quickly to the stationary distribution. It has been shown that the Glauber dynamics converge exponentially fast to the stationary distribution in KL divergence if the model satisfies a *Modified Log-Sobolev Inequality (MLSI)*, a functional inequality that is weaker than the usual Log-Sobolev inequality in discrete spaces Bobkov and Tetali (2006). The MLSI has been established for a variety of Ising models under different constraints Marton (2015); Anari et al. (2021); Chen and Eldan (2022); Blanca et al. (2022); Anari et al. (2024b).

Our first main Theorem establishes estimation guarantees for Ising models of bounded operator norm that satisfy MLSI, when running the MPLE over some set of interaction matrices. Crucially, the only properties that this set needs to satisfy are that J^* belongs to a set of matrices of bounded operator norm. This flexibility with respect to the optimizing set enables us to obtain polynomial-time algorithms in various cases, particularly when the set is convex and admits efficient projections (see Section 2.1.2 for more discussion). The estimation guarantees are phrased in terms of Frobenius norm closeness to the matrix J^*, but we will see in Sections 2.1 and 2.1.2 how these can be easily translated to bounds on the TV distance. Also, in the formal version of the theorem, we have included the case of a non-zero external field, which doesn't change the analysis in any significant way and is omitted here for simplicity.

†. The $\widetilde{O}(n^8/\epsilon^4)$ sample complexity can be explicitly derived from Theorem 7.3 of Klivans and Meka (2017). However, by combining the Frobenius-norm learning result of Dagan et al. (2021) with arguments similar to those in Klivans and Meka (2017), one can derive $\widetilde{O}(n^4/\epsilon^2)$ sample complexity for MPLE, similar to our result.

Theorem 1 (informal, see Theorem 28) *Suppose we are given l independent samples $X^{(1)}, \dots, X^{(l)} \sim \mathbf{Pr}_{J^*}$ and $\mathbf{Pr}_{J^*}$ satisfies MLSI. Let $\mathcal{R} \subseteq \mathcal{S}_0^n$ be a subset of matrices such that $\sup_{A \in \mathcal{R}} \|A\|_2 = O(1)$ and $J^* \in \mathcal{R}$. Then, running MPLE with optimizing set $\mathcal{R}$ produces an estimate $\hat{J} \in \mathcal{R}$, such that with high probability over the choice of samples $\|J^* - \hat{J}\|_F \leq \epsilon$, as long as $l = \tilde{\Omega}(n^2/\epsilon^2)$.*

As far as we know, a result of this generality has not appeared in the literature. The most related prior result is Theorem 5.2 from Anari et al. (2024a), where they obtain an elegant result for TV estimation using MPLE for Ising models satisfying *Approximate Tensorization of Entropy (ATE)*, which is a stronger functional inequality than MLSI. Moreover, they require that every matrix in the optimizing set of MPLE satisfies ATE, which places strong constraints on the choice of this set. Finally, they require the matrices to be of *bounded-width*, which is a stronger assumption than bounded operator norm and results in the loss of additional polynomial factors. We remark that the sample complexity in Theorem 1 has an exponential dependence on the MLSI constant and this is reflected in the applications of this result, while the prior result is free of such dependence (see Theorem 28 for the precise dependence). We next present a variety of applications of the main result for TV learning in classes of Ising models of bounded operator norm.

2.1.1. APPLICATION: ESTIMATING SPECTRALLY-BOUNDED ISING MODELS IN TV

Perhaps the most widely-studied condition that enables computationally efficient sampling and inference in Ising models is *Dobrushin's Uniqueness Condition* Dobruschin (1968), which asserts that $\|J\|_\infty < 1$, or equivalently $\sum_j |J_{ij}| < 1$ for all rows i. This condition has been shown to imply a variety of structural properties for the Ising measure, such as fast mixing Levin and Peres (2017), correlation decay Künsch (1982), and concentration inequalities Götze et al. (2021); Adamczak et al. (2019); Marton (2015).

Unfortunately, this condition is sometimes too strict and does not capture the tractable regime of an Ising model. A notable example is the celebrated *Sherrington-Kirkpatric (SK)* model, where J^* is a random matrix with each J^*_{ij} sampled independently from $\mathcal{N}(0, \beta^2/n)$, where $\beta > 0$ is a parameter called the *inverse temperature* of the model. Standard random matrix theory arguments Anderson et al. (2010) can be used to show that the operator norm is bounded if β is bounded. In contrast, it is straightforward to observe that the expected l_1-norm of each column is $\Theta(\beta\sqrt{n})$, thus Dobrushin's condition is only satisfied if $\beta = O(1/\sqrt{n})$. However, it is expected that the model exhibits weak interactions for all sufficiently small constant β.

Motivated by this gap, Eldan et al. (2022) introduced an alternative condition for fast mixing. In particular, we say that an Ising model as in (1) is *spectrally bounded* if $\lambda_{max}(J^*) - \lambda_{min}(J^*) < 1$(note that J^* is symmetric, hence diagonalizable). In Anari et al. (2021), they prove that if a model is spectrally bounded, then MLSI holds and the Glauber dynamics mix in polynomial time. Thus, we can apply Theorem 1 for this class of Ising models, which results in information-theoretically optimal sample complexity $\tilde{O}(n^2/\epsilon^2)$ for estimating spectrally bounded Ising models in TV.

Corollary 2 (informal, see Corollary 44) *Suppose we are given l independent samples $X^{(1)}, \dots, X^{(l)} \sim \mathbf{Pr}_{J^*}$, where $\lambda_{max}(J^*) - \lambda_{min}(J^*) < 1 - \alpha$ and J^* has zero-diagonal. Then, there is a polynomial time algorithm (MPLE) that produces an estimate $\hat{J} \in \mathcal{S}_0^n$, such that with high probability over the choice of samples we have $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \leq \epsilon$, as long as $l = \tilde{\Omega}(\exp(1/\alpha^{5/4}) \cdot n^2/\epsilon^2)$.*

The implicit constant in the sample complexity contains additional sub-polynomial factors of the form $e^{\sqrt{\log n}}$. As far as we know, the most relevant prior work is Anari et al. (2024a), where

they prove that MPLE succeeds in finding a model that is ϵ-close to the true Ising model $\mathbf{Pr}_{J^*}$ using $O(n^{3+C}/\epsilon^4)$ samples for some $C < 1$, by establishing ATE with an inverse polynomial constant. Subsequent work Lee (2023) has shown that, in fact, ATE holds with a $\Theta(1)$ constant in this setting, which can be used to remove the C from the exponent, yielding $O(n^3/\epsilon^4)$ sample complexity. Our result thus improves over this bound in terms of the dependence on n, ϵ, by showing that the MPLE actually achieves the information theoretically optimal sample complexity $O(n^2/\epsilon^2)$ for estimating Ising models in TV Devroye et al. (2020). We should remark, though, that the implicit constant in the sample complexity of Corollary 2 contains a factor that is exponential in $1/\alpha^2$, while the bound in Anari et al. (2021) is free of such dependence. As noted above, spectrally bounded models do not necessarily have bounded width (see Section 2.2 for definition) e.g. for the SK model, $\|J\|_\infty$ could be $\Theta(\sqrt{n})$, so the prior work Klivans and Meka (2017) would give exponential sample complexity.

2.1.2. Application: Estimating the SK-model in TV

As mentioned in Section 2.1, the SK model is one of the canonical examples of a mean-field model in statistical physics, exhibiting fascinating phase transition phenomena that have been the subject of extensive study in probability theory Panchenko (2012); Talagrand (2010). The relevant parameter is the inverse temperature $\beta > 0$. Standard results from random matrix theory imply that if $\beta < 1/4$, then with high probability the interaction matrix has spectrum inside an interval of size < 1, which means the model is spectrally bounded. Thus, in this regime, Corollary 2 can be used to learn the model optimally with $\tilde{O}(n^2/\epsilon^2)$ samples.

However, it turns out that efficiently learning the model in TV distance is possible for much larger values of β. In particular, in Gaitonde and Mossel (2024), the authors remarkably prove that a polynomial time algorithm introduced in Wu et al. (2019) actually estimates the SK model in TV as long as $\beta < 1$. In a subsequent work, Chandrasekaran and Klivans (2024) shows that the same algorithm succeeds even when $\beta = O(\sqrt{n})$, which extends well into the low-temperature region of the model.

While these works greatly push the frontiers of efficient learnability, the sample complexity arising from these results is of the order of $\tilde{O}(n^9/\epsilon^8)$. In the next Corollary, we use Theorem 1 and recently established MLSI in Anari et al. (2024b) to obtain $\tilde{O}(n^4/\epsilon^2)$ sample complexity for learning the SK-model up to $\beta \approx 0.295$, which is beyond the threshold of spectrally-bounded models.

Corollary 3 (informal, see Corollary 42) *Suppose we are given l independent samples $X^{(1)}, \ldots, X^{(l)} \sim \mathbf{Pr}_{J^*}$, where J^* is sampled according to the SK-model with $\beta < C$, where $C \approx 0.295$. Then, there is a polynomial time algorithm (MPLE) that produces an estimate $\hat{J} \in \mathcal{S}_0^n$, such that with high probability over the choice of samples and the choice of matrix J^* we have $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \leq \epsilon$, as long as $l = \tilde{\Omega}(n^4/\epsilon^2)$.*

While our result can only accommodate a short range of values of β beyond high temperature, compared to the range $\beta = O(\sqrt{\log n})$ of Chandrasekaran and Klivans (2024), it greatly improves the sample complexity. An interesting avenue for future work would be to determine whether this could be further improved to match the optimal sample complexity $O(n^2/\epsilon^2)$ or whether a computational-statistical gap exists beyond high temperature.

Closely related to the SK-model are *diluted* versions, where the matrix is supported on a sparse graph. One such version, which we call for simplicity the *diluted SK-model*, arises from sampling

a random d-regular graph G, where the matrix J^* will be supported. Every non-zero entry of J^* is then sampled independently and uniformly from $\{-\beta/\sqrt{d}, \beta/\sqrt{d}\}$. Standard results from random matrix theory again imply that if $\beta < 0.25$, then the model is spectrally bounded with high probability and $\tilde{O}(n^2/\epsilon^2)$ samples suffice by Corollary 2. Chandrasekaran and Klivans (2024) show that TV learning in polynomial time is possible if $\beta = O(\sqrt{\log n})$, with $\tilde{O}(n^8 d/\epsilon^8)$ samples. We use the recently established MLSI for diluted SK up to $\beta \approx 0.295$ Anari et al. (2024b) to establish that $\tilde{O}(n^4/\epsilon^2)$ samples suffice in that regime if we run MPLE.

Corollary 4 (informal, see Corollary 43) *Suppose we are given l independent samples $X^{(1)}, \dots, X^{(l)} \sim \mathbf{Pr}_{J^*}$, where J^* is sampled according to the diluted SK-model with $\beta < C$, where $C \approx 0.295$. Then, there is a polynomial time algorithm (MPLE) that produces an estimate $\hat{J} \in \mathcal{S}_0^n$, such that with high probability over the choice of samples and the choice of matrix J^* we have $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \leq \epsilon$, as long as $l = \tilde{\Omega}(n^4/\epsilon^2)$.*

Note that in this setting $\|J^*\|_\infty = O(\sqrt{d})$, so one could use the result of Dagan et al. (2021) about learning in Frobenius norm together with Lemma 39 that connects TV and Frobenius norms to prove that $O(e^{\Theta(\sqrt{d})} n^4)$ samples suffice for TV learning using MPLE. However, notice that even if the degree grows mildly with the number of nodes, i.e., $d = \omega(\log n)$, the sample complexity suffers from additional polynomial factors (or worse). In contrast, the sample complexity in Corollary 43 only contains a sub-polynomial $\exp(\sqrt{\log n})$ factor, regardless of the value of d.

2.1.3. APPLICATION: ANTIFERROMAGNETIC EXPANDERS

Another prominent class of models is the ones where there is a gap between the largest and second-largest eigenvalue of the adjacency matrix. When the model is antiferromagnetic, then the spectrum essentially consists of a very negative eigenvalue and a bulk that is concentrated on a small interval. Prior work Anari et al. (2024b); Koehler et al. (2022b) has shown that one can "ignore" this very negative eigenvalue and establish MLSI in this case. Thus, if we can efficiently project on this set of matrices, then MPLE runs in polynomial time and has the optimal sample complexity. We show that this is indeed possible, which gives rise to the following Corollary.

Corollary 5 (informal, see Corollary 45) *Let $\alpha \in (0,1), c > 0$ be constants and $\mathbf{1}$ the all-ones vector. Define the set $\mathcal{R} \subseteq \mathcal{S}_0^n$ of matrices that have $\mathbf{1}$ as an eigenvector with eigenvalue $-c$ and the rest of the spectrum is on an interval of size α around 0. Suppose $J^* \in \mathcal{R}$. Then, given l independent samples from $\mathbf{Pr}_{J^*}$, the MPLE over $\mathcal{R}$ can be implemented in polynomial time and returns $\hat{J}$ such that $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \leq \epsilon$ with high probability, as long as $l = \tilde{\Omega}(n^2/\epsilon^2)$.*

For a canonical example in this set, consider the adjacency matrix A_G of a random d-regular graph G and take $J^* = -\beta A_G$. Then, from Friedman (2003) if follows that J^* belongs in the set $\mathcal{R}$ with $c = \beta d$ and $\alpha = 4\beta\sqrt{d-1}$ when we take $\beta < 1/(4\sqrt{d-1})$. Thus, we can learn this model in TV distance optimally and efficiently. The most relevant prior work in this case is Koehler et al. (2024), which covers this class of models since it allows some eigenvalues to be very negative. Yet, the sample complexity is $\tilde{O}(n^4 R^2/\epsilon^4)$, where R is the width of the model, which could be $\Theta(\sqrt{n})$ in that case (see Remark 46 for an example). Since $R = \Theta(\sqrt{n})$ in the worst case, the bounded width result of Dagan et al. (2021) does not apply.

2.2. Estimating Bounded-Width Ising Models in TV

We say that an Ising model has *bounded width*, if the interaction matrix is assumed to have infinity norm bounded by some constant $M > 0$, i.e. $\|A\|_\infty \leq M$. We know that if $\|A\|_\infty \leq M$, then $\|A\|_2 \leq M$ and thus $A \in \mathcal{R}$. Hence, this set of matrices is a subset of $\mathcal{R}$ that was considered in Theorem 1. However, note that M could be an arbitrary constant, which means the model could exhibit long-range correlations, Glauber dynamics might mix exponentially slowly (see e.g. Mossel et al. (2009)), and concentration of measure in general fails to hold. Our first contribution involves an improved analysis of the MPLE estimator, which results in the following guarantee for estimating the model $\mathbf{Pr}_{J^*}$.

Theorem 6 (informal, see Theorem 37) *Suppose we are given l independent samples $X^{(1)}, \ldots, X^{(l)} \sim \mathbf{Pr}_{J^*}$, where $\|J^*\|_\infty \leq M$. Then, if $\hat{J}$ is the MPLE estimator, with high probability over the choice of samples we have, as long as $l = \tilde{\Omega}(n^2/\epsilon)$,*

$$\mathop{\mathbf{E}}_{X \sim \mathbf{Pr}_{J^*}} [\|(\hat{J} - J^*)X\|_2^2] \leq \epsilon. \tag{2}$$

The implicit constant in the bound above contains a factor $\exp(M)$. The guarantee provided by Theorem 6 might seem non-standard, but we will see that it is well-suited for estimation in TV distance in the following section.

2.2.1. Applications

First, as a direct corollary of Theorem 6 (proved in Section F), we can obtain the following.

Corollary 7 (informal) *Suppose we are in the setting of Theorem 6. Then, with high probability over the choice of samples, we have, as long as $l = \tilde{\Omega}(n^2/\epsilon)$,*

$$\|\hat{J} - J^*\|_F^2 \leq \epsilon. \tag{3}$$

Corollary 7 also appears as Corollary 6 in Dagan et al. (2021), hence we recover the previously established guarantees for learning in the Frobenius norm. Note that in general, if J^* is in low temperature, $\mathbf{E}_{J^*}[\|(\hat{J} - J^*)X\|_2^2]$ could be significantly larger than $\|\hat{J} - J^*\|_F^2$ (we also give such examples in Section F), so Theorem 6 is a strict improvement over the result of Dagan et al. (2021).

Now we are ready to state the implications of our results for learning in TV. First, we note that without imposing any additional assumptions, we can obtain a sample complexity of $\tilde{O}(n^4)$ from the Frobenius norm approximation. The reason is that one can show that an $O(\epsilon)$ approximation in Frobenius norm implies an $O(n\epsilon)$ approximation in TV, using similar arguments to Theorem 7.3 in Klivans and Meka (2017). For completeness, we give a self-contained proof of this fact in Section E. Thus, the following result follows from this connection together with Corollary 7.

Corollary 8 (informal) *Suppose we are in the setting of Theorem 6. Then, if $l = \Omega(n^4/\epsilon^2)$, with high probability over the choice of sample, $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \leq \epsilon$.*

We now show how we can improve on the $O(n^4)$ sample complexity of Corollary 8 using the refined analysis of Theorem 6. To do that, we will assume that the second moments of the true model are "robust" to small perturbations of the matrix. Intuitively, we expect this to happen whenever J^* is away from the critical temperature where a phase transition occurs. Formally, let us introduce the following regularity assumption.

Definition 9 *We say an Ising Model $\mathbf{Pr}_{J^*}$ satisfies (γ, C)-regularity for some $\gamma, C > 0$ if the following holds: for any $J \in \mathcal{S}_0^n$ such that $\mathbf{E}_{J^*}[\|(J-J^*)X\|_2^2] \leq \gamma$, we have $\mathbf{E}_J[\|(J-J^*)X\|_2^2] \leq C \cdot \mathbf{E}_{J^*}[\|(J-J^*)X\|_2^2]$.*

Of course, the crucial part of this definition is the scaling relation between γ and $C > 0$. We show as a Corollary of Theorem 6 that if a model is $(1/n, 1)$-regular, then $O(n^3)$ samples suffice for TV learning.

Corollary 10 (informal, see Corollary 38) *Suppose we are in the setting of Theorem 6. Additionally, assume there exist constants $\gamma, C > 0$ such that $\mathbf{Pr}_{J^*}$ satisfies $(\gamma/n, C)$-regularity. Then, for any $\epsilon > 0$, if $l = \Omega(n^3/\epsilon^2)$, with high probability over the choice of samples $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \leq \epsilon$.*

Note that the sample complexity of Corollary 10 is only a factor $O(n)$ away from the optimal sample complexity of Devroye et al. (2020). This regularity condition covers a wide range of models that do not necessarily need to be in high temperature. In particular, in Section G, we prove that the Curie-Weiss model in low temperature satisfies the condition. Of course, we can also show that models satisfying more familiar conditions such as Dobrushin's condition and spectrally-bounded models also satisfy this regularity condition (see Section G).

3. Related Work

Learning MRFs from multiple samples. The problem of estimating a Markov Random Field (MRF) from multiple independent samples from the model has a rich history, starting from the seminal work Chow and Liu (1968) from the 1960s, showing that for the Ising model, if the graph structure of the model is a tree, then Maximum Likelihood Estimation (MLE) can be solved in polynomial time. Information theoretically, Devroye et al. (2020) establishes the minimax rate for estimating Ising models in TV as $\Theta(|E|/\epsilon^2)$, where $|E|$ is the number of non-zero entries of the interaction matrix (see also Brustle et al. (2020) for an alternative argument using linear programming). For the task of estimating the structure of Ising models with arbitrary graph topology and bounded degree d, the breakthrough work of Bresler (2015) provided the first polynomial time algorithm, where the sample complexity is doubly exponential in d. Subsequent works Hamilton et al. (2017); Vuffray et al. (2016); Klivans and Meka (2017); Wu et al. (2019) proposed new algorithms with improved guarantees. In particular, Klivans and Meka (2017) obtains the first polynomial time algorithm for learning the structure of bounded-width models, while only requiring $O(e^d \log n)$ independent samples using l_2-regularized per-node logistic regression. This matched the information-theoretic lower bound from Santhanam and Wainwright (2012). In the case of latent variables, Bresler et al. (2019) gives a polynomial-time algorithm for learning ferromagnetic Restricted Boltzmann Machines. Beyond bounded width, a recent line of work studies Ising models under spectral constraints on the interaction matrix. Anari et al. (2024a) shows that for spectally bounded models, MPLE succeeds in TV learning with $O(n^{3+C})$ samples for some constant C. For the SK model, this implies efficient learning for all $\beta < 1/4$. Gaitonde and Mossel (2024) is the first work to obtain a polynomial time algorithm for learning the SK-model all the way up to $\beta < 1$, and Chandrasekaran and Klivans (2024) extends the range of efficient learning for all $\beta = o(\sqrt{\log n})$.

A related line of work studies the problem of learning the structure of MRFs using samples from the trajectory of Glauber dynamics Bresler et al. (2017); Gaitonde and Mossel (2024); Gaitonde et al. (2025). The recent work of Gaitonde et al. (2024) provides a near-linear time algorithm that

learns the structure of a t-th order MRF using $O(n \log n)$ updates of Glauber dynamics, bypassing fundamental barriers for efficient higher order MRF estimation from independent samples. The work Jayakumar et al. (2024), which studies learning from independent samples of a metastable state of the Glauber dynamics, is also close in spirit.

The particular case of MRFs with tree structure has also received attention, starting with Chow and Liu (1968). Bhattacharyya et al. (2021); Daskalakis and Pan (2021) establish that the Chow-Liu algorithm is information-theoretically optimal for finite samples. For the related problem of estimating the low-dimensional marginals of the model, Bresler and Karzand (2020); Boix-Adsera et al. (2022) give polynomial-time and sample-optimal algorithms, and Nikolakakis et al. (2021) studies the setting where noisy labels of the nodes are observed. Finally, Kandiros et al. (2023) provides guarantees for polynomial-time estimation of latent tree Ising models in TV.

Finally, a recent line of work aims at estimating Ising models using score matching Koehler et al. (2022a); Koehler and Vuong (2023); Koehler et al. (2024). In particular, as noted in Koehler et al. (2024), they focus on a class of low-complexity Ising models. Remarkably, they obtain sample complexity bounds that scale polynomially with the width of the model, in contrast to the exponential dependence in most prior work on MRF estimation.

Learning MRFs from a single sample. In this line of work, it is usually assumed that the true model belongs in a class of "low-dimensional" models and the task is to estimate it given a *single* n-dimensional observation from the model. In the case of an Ising model whose interaction matrix is known up to a scalar parameter β, Chatterjee (2007) initially showed that MPLE is $\sqrt{n}$-consistent for β using the technique of exchangeable pairs. Bhattacharya and Mukherjee (2018) extended these results under general conditions on the log-partition function. Ghosal and Mukherjee (2020) studied the problem when, in addition to β, there is also an unknown scalar parameter in the external field, and Daskalakis et al. (2019) generalized the result for logistic regression with dependencies. Dagan et al. (2021) provided estimation guarantees when the interaction matrix lies in a low-dimensional subspace. Several variations of these settings have been studied, including optimal joint estimation of parameters for logistic regression with dependencies Kandiros et al. (2021); Mukherjee et al. (2021), estimation of tensor Ising models Daskalakis et al. (2020); Mukherjee et al. (2022), estimation of hard-constrained models Bhattacharya and Ramanan (2021); Galanis et al. (2024a), inference on dense graphs Xu and Mukherjee (2023).

Sampling Ising models. There is a vast literature in probability theory that focuses on proving fast sampling of Ising models under different constraints. Here, we focus on reviewing the results that are most relevant to the classes of Ising models that we study. Modified Log-Sobolev inequalities were introduced in Bobkov and Tetali (2006) to prove fast mixing of Markov Chains in discrete spaces. The classical Dobrushin's condition has been known to imply MLSI Levin and Peres (2017) and is tight in the case of the Curie-Weiss model. The class of spectrally bounded models, where $\lambda_{\max}(J) - \lambda_{\min}(J) < 1$, was introduced in Eldan et al. (2022) to capture relevant models from statistical physics, such as the SK-model. They established the Poincaré inequality when $\beta < 1/4$, complementing the result of Bauerschmidt and Bodineau (2019) that proved a version of the log-Sobolev inequality in the same regime. Subsequently, Anari et al. (2021) established the MLSI for spectrally bounded models, which implies optimal mixing of the Glauber dynamics. A different proof using localization schemes was given in Chen and Eldan (2022). In the work Lee (2023), they establish the stronger ATE property for these models. The condition of spectrally bounded models

was shown to be tight for polynomial time sampling in Kunisky (2024); Galanis et al. (2024b). A polynomial time algorithm for sampling from the SK model in Wasserstein distance using algorithmic stochastic localization was given for $\beta < 1/2$ in El Alaoui et al. (2022) and extended to all $\beta < 1$ in Celentano (2024). The recent work of Anari et al. (2024b) established an MLSI for the SK-model for $\beta < 0.295$. Finally, the class of bounded-width Ising models includes many examples where sampling from the model is NP-hard. Indeed, for d-regular graphs with $\beta > \beta_c$, where β_c is the Kesten-Stigum threshold, Sly (2010); Sly and Sun (2012); Galanis et al. (2015) show that approximate sampling from the distribution is NP-hard.

4. Technical Contributions

We first describe the algorithm that is employed for all results. The most common approach for obtaining an estimate $\hat{J}$ is to compute the *Maximum Likelihood Estimator (MLE)* given the samples. In the case of Ising models, to compute the MLE, one has to calculate the probability of the observed samples under different models. However, this involves computing the partition function of the model Z_J, which is NP-hard even to approximate in many interesting regimes Sly (2010); Sly and Sun (2012); Galanis et al. (2015, 2024b).

An attractive alternative, first proposed in Besag (1974), is the so-called *Maximum Pseudo-Likelihood Estimator (MPLE)*. Computing the Pseudo-Likelihood of a given model involves computing the conditional probability of the spin of a node i conditioned on the spins of all the other nodes. Formally, suppose we get independent samples $X^{(1)}, \dots, X^{(l)} \sim \mathbf{Pr}_{J^*}$. Then, the MPLE over a set of matrices $\mathcal{R} \subseteq \mathcal{S}_0^n$ is defined as

$$\hat{J} := \arg\max_{J \in \mathcal{R}} PL(J; X^{(1)}, \dots, X^{(l)}) := \arg\max_{J \in \mathcal{R}} \prod_{k=1}^{l} \prod_{i=1}^{n} \mathbf{Pr}_J[X_i^{(k)} | X_{-i}^{(k)}] \tag{4}$$

The set $\mathcal{R}$ will be chosen depending on the particular class of Ising models we are trying to estimate. One useful property of this objective function is that it is a concave function of J, which makes it easy to optimize using first-order methods whenever $\mathcal{R}$ is a convex set that admits efficient projections. To make calculations more convenient, we will consider instead minimizing the negative log pseudolikelihood, which we call ϕ. The advantage is that the objective in MPLE has a simple closed form that does not involve the partition function, which is why it is preferred over MLE. Since the optimization takes place in a high-dimensional space, we will be interested in computing the first and second derivatives of ϕ at a point $J \in \mathcal{S}_0^n$ and at direction $A \in \mathcal{S}_0^n$. These are given by the formulas (see also Section(A))

$$\frac{\partial \phi(J^*)}{\phi A} = \sum_{k=1}^{l} \sum_{i=1}^{n} (A_i X^{(k)})(\tanh(J_i X^{(k)}) - X_i^{(k)}), \quad \frac{\partial^2 \phi(J)}{\partial A^2} = \sum_{k=1}^{l} \sum_{i=1}^{n} (A_i X^{(k)})^2 \operatorname{sech}(J_i X^{(k)})^2 \tag{5}$$

The standard way of analyzing the MPLE in the single sample literature Chatterjee (2007); Bhattacharya and Mukherjee (2018); Daskalakis et al. (2019); Dagan et al. (2021) is to upper bound the first derivative at J^* for all directions A and to lower bound the second derivative for all J and for all directions A. This, combined with a union bound argument, suffices for obtaining estimation guarantees for $\hat{J}$. Since these derivatives are random quantities that depend on the samples, to bound them, we introduce a number of technical novelties and combine a variety of tools from high-dimensional probability. We now highlight these contributions in the two settings we study.

4.1. Bounded Operator Norm Models

The work that is closest in spirit to the level of generality we are aiming for is Dagan et al. (2021), where they analyze the MPLE when the infinity norm of the matrix is $O(1)$(bounded-width). In our case, we only know that the operator norm is bounded, which allows for potentially unbounded infinity norm (as is often in applications, e.g. the SK model). As we shall see, establishing tight upper and lower bounds for the first and second derivatives of the pseudolikelihood function ϕ when the infinity norm is unbounded poses additional challenges, which we now describe.

The first step towards that goal is to establish upper bounds for the first derivative of the pseudolikelihood function $\partial\phi(J^*)/\partial A$ for all directions $A \in \mathcal{R}$, where $\mathcal{R}$ is the optimizing set. In Dagan et al. (2021), the authors rely on concentration bounds which hold assuming the Approximate Tensorization of Entropy (ATE) inequality holds, which is a stronger functional inequality than MLSI. Instead, we show that MLSI suffices by using a general result about two-level concentration from Sambale and Sinulis (2019). In the proof, we show that even if we only know that the operator norm of the matrix A is bounded, the concentration radius from Sambale and Sinulis (2019) can be bounded by $O(l \cdot \|A\|_F^2)$. Since the expectation of the first derivative is 0 when evaluated by J^*, this implies that with high probability $O|\partial\phi(J^*)/\partial A| = O(l \cdot \|A\|_F^2)$.

The second and most crucial step in the analysis of MPLE is to lower bound the second derivative $\partial^2\phi(J)/\partial A^2$ of the pseudolikelihood function for all J and all directions A. Ignoring the term involving sech momentarily, which comes with additional challenges, the second derivative in (5) is a second-degree polynomial of the Ising model. Using again the machinery from Sambale and Sinulis (2019), we can show that this polynomial concentrates at an $O(l \cdot \|A\|_F^2)$ radius with high probability. To conclude that the second derivative is large, we need to establish that, on expectation, this polynomial is $\Omega(l \cdot \|A\|_F^2)$. This lower bound would certainly hold if we had a product distribution instead of an Ising model. Motivated by that, we use the well-known Hubbard-Stratonovich transform to decompose the model into a mixture of product distributions with external fields. The distribution of the external fields in this mixture is well known, and we use it to lower bound the expectation of this polynomial for the majority of these product measures, which suffices to obtain the desired lower bound on the expectation.

Finally, if we wish to lower bound the second derivative, we have to lower bound the term involving sech in (5). If the infinity norm of the matrices was bounded, as in Dagan et al. (2021), then this step is trivial, as $|J_i X^{(k)}| = O(1)$ always. In contrast, with an operator norm bound, this term could be as large as $\Theta(\sqrt{n})$, which would result in an exponentially small lower bound for the second derivative in the worst case. Our solution is to write $J_i X^{(k)} = J_i^* X^{(k)} + (J_i - J_i^*)X^{(k)}$. Since J^* is fixed, we can use the previous concentration results to bound $J_i^* X^{(k)}$ with high probability. For $(J_i - J_i^*)X^{(k)}$, we can also use the preceding concentration results to bound it by $\|J - J^*\|_F$ with high probability. However, since we are considering an arbitrary matrix J, we cannot know *a priori* that $\|J - J^*\|_F$ will be small. Intuitively, J is any matrix we wish to show satisfies $\phi(J) > \phi(J^*)$. To address this issue, we instead focus the analysis of the second derivative only on a *ring* of matrices

$$\mathcal{R}_\epsilon := \{J \in \mathcal{R} : \epsilon \leq \|J - J^*\|_F \leq 2\epsilon\}$$

The reason we take a ring is that on the one hand we would like $\|J - J^*\|_F$ to be upper bounded in order to lower bound the sech term, but on the other hand we would like $\|J - J^*\|_F$ to be lower bounded, so that we are guaranteed that with high probability the second derivative dominates the

first. Indeed, using all the previous observations and some careful union bound arguments over a suitable discrete net of points, we can establish that with high probability $\forall J \in \mathcal{R}_\epsilon$ we have $\phi(J) > \phi(J^*)$. It turns out that this also implies that $\phi(J) > \phi(J^*)$ for all J with $\|J - J^*\|_F > 2\epsilon$, by using the convexity of ϕ. Thus, we can conclude that $\|\hat{J} - J^*\|_F$ will be small.

4.2. Bounded Width Models

In Klivans and Meka (2017), they obtain TV learning guarantees by using the guarantees for structure learning, which ensures that $\max_{ij} |J_{ij} - J^*_{ij}| \leq \epsilon$ using $O(\log n/\epsilon^4)$ samples. They subsequently set the accuracy to ϵ/n^2, so that it guarantees closeness in TV. However, requiring accuracy for each entry of J^* is a very strict condition, which results in sample complexity $O(n^8)$. In this work, we instead obtain guarantees of closeness to J^* through different metrics, which, while counterintuitive at first, are more suitable for implying closeness in TV.

We first discuss how to get the improved guarantee of Theorem 6. Since bounded-width models could be in low temperature, we cannot, in general, rely on concentration bounds for the original distribution $\mathbf{Pr}_{J^*}$. We follow the strategy presented in Dagan et al. (2021), which involves finding a small ($O(\log n)$) collection of subsets $I_j \subseteq [n]$, such that conditioned on each subset the model satisfies Dobrushin's condition, and every node $i \in [n]$ is contained in a constant fraction of the subsets. For the second derivative, we notice that it can be lower bounded by a larger quantity than that in Dagan et al. (2021), without breaking it into parts, by simply conditioning on one of the subsets I_j. Indeed, using (5) and the bounded-width property, we can lower bound the conditional expectation of the second derivative by the variance of a linear function of the model as follows.

$$\begin{aligned}\mathbf{E}\left[\frac{\partial^2\phi(J)}{\partial A^2}\middle| X_{-I_j}\right] &\geq \operatorname{sech}(M)^2 \sum_{k=1}^{l}\sum_{i=1}^{n} \mathbf{E}\left[\left(A_i X^{(k)}\right)^2 \middle| X^{(k)}_{-I_j}\right] \\ &= \operatorname{sech}(M)^2 \sum_{k=1}^{l}\sum_{i=1}^{n} \left(\mathbf{Var}\left(A_i X^{(k)} \middle| X^{(k)}_{-I_j}\right) + \mathbf{E}\left[A_i X^{(k)} \middle| X^{(k)}_{-I_j}\right]^2\right)\end{aligned}$$

In the above, we have used the definition of the conditional variance. We can show that the conditional variance summed over all nodes is of the order $\Theta(\|A_{I_j}\|_F^2)$, since the conditional model satisfies Dobrushin's condition. For the variance of the conditional expectation, there is no general formula, so we would like to approximate it by its expectation $\mathbf{E}[\|E[AX|X_{-I_j}]\|_2^2]$. Simply using the Chernoff bound for the independent samples does not suffice, because we would like this lower bound to hold uniformly for all directions A, and the union bound will incur additional polynomial factors. We avoid this union bounding argument by a careful application of matrix Bernstein's inequality. This enables us to approximate the second term in the sum by $\mathbf{E}[\|E[AX|X_{-I_j}]\|_2^2]$ uniformly over all matrices. These insights result in a lower bound of $\Omega(l \cdot \mathbf{E}_{J^*}[\|AX\|_2^2])$ for the second derivative.

For the first derivative, we can upper bound it by $O(l \cdot \mathbf{E}_{J^*}[\|AX\|_2^2])$ uniformly for all matrices A by using the technique of splitting it into parts and a similar application of matrix Bernstein's inequality. Since both the first and second derivatives are upper and lower bounded by the same quantity, we can proceed as previously using a Taylor expansion for ϕ and establish that $E_{J^*}[\|(\hat{J} - J^*)X\|_2^2] \leq \epsilon$ with high probability.

Acknowledgement

Constantinos Daskalakis and Rui Yao were supported by a Simons Investigator Award, a Simons Collaboration on Algorithmic Fairness, ONR MURI grant N00014-25-1-2116, ONR grant N00014-25-1-2296. Vardis Kandiros was supported by a Postdoctoral Fellowship of the Data Science Institute at Columbia University.

References

Radosław Adamczak, Michał Kotowski, Bartłomiej Polaczyk, and Michał Strzelecki. A note on concentration for polynomials in the ising model. 2019.

Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence i: Modified log-sobolev inequalities for fractionally log-concave distributions and high-temperature ising models. *arXiv preprint arXiv:2106.04105*, 2021.

Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Universality of spectral independence with applications to fast mixing in spin glasses. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5029–5056. SIAM, 2024a.

Nima Anari, Frederic Koehler, and Thuy-Duong Vuong. Trickle-down in localization schemes and applications. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1094–1105, 2024b.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Number 118. Cambridge university press, 2010.

Shiri Artstein-Avidan, Apostolos Giannopoulos, and Vitali D Milman. *Asymptotic geometric analysis, Part II*, volume 261. American Mathematical Society, 2021.

Roland Bauerschmidt and Thierry Bodineau. A very simple proof of the lsi for high temperature spin systems. *Journal of Functional Analysis*, 276(8):2582–2588, 2019.

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

Bhaswar B Bhattacharya and Sumit Mukherjee. Inference in ising models. 2018.

Bhaswar B Bhattacharya and Kavita Ramanan. Parameter estimation for undirected graphical models with hard constraints. *IEEE Transactions on Information Theory*, 67(10):6790–6809, 2021.

Arnab Bhattacharyya, Sutanu Gayen, Eric Price, and NV Vinodchandran. Near-optimal learning of tree-structured distributions by chow-liu. In *Proceedings of the 53rd annual acm SIGACT symposium on theory of computing*, 2021.

Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda. On mixing of markov chains: Coupling, spectral independence, and entropy factorization. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3670–3692. SIAM, 2022.

Sergej G Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

Sergey G Bobkov and Prasad Tetali. Modified logarithmic sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19(2):289–336, 2006.

Enric Boix-Adsera, Guy Bresler, and Frederic Koehler. Chow-liu++: Optimal prediction-centric learning of tree ising models. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 417–426. IEEE, 2022.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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

Guy Bresler and Mina Karzand. Learning a tree-structured ising model in order to make predictions. *The Annals of Statistics*, 48(2):713–737, 2020.

Guy Bresler, David Gamarnik, and Devavrat Shah. Learning graphical models from the glauber dynamics. *IEEE Transactions on Information Theory*, 64(6):4072–4080, 2017.

Guy Bresler, Frederic Koehler, and Ankur Moitra. Learning restricted boltzmann machines via influence maximization. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 828–839, 2019.

Johannes Brustle, Yang Cai, and Constantinos Daskalakis. Multi-item mechanisms without item-independence: Learnability via robustness. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 715–761, 2020.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Pietro Caputo. Lecture notes on entropy and markov chains. *Preprint, available from: http://www. mat. uniroma3. it/users/caputo/entropy. pdf*, 2023.

Pietro Caputo, Georg Menz, and Prasad Tetali. Approximate tensorization of entropy at high temperature. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 24, pages 691–716, 2015.

Michael Celentano. Sudakov–fernique post-amp, and a new proof of the local convexity of the tap free energy. *The Annals of Probability*, 52(3):923–954, 2024.

Gautam Chandrasekaran and Adam Klivans. Learning the sherrington-kirkpatrick model even at low temperature. *arXiv preprint arXiv:2411.11174*, 2024.

Sourav Chatterjee. Estimation in spin glasses: A first step. 2007.

Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 110–122. IEEE, 2022.

Zongchen Chen, Kuikui Liu, and Eric Vigoda. Optimal mixing of glauber dynamics: Entropy factorization via high-dimensional expansion. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1537–1550, 2021.

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

Mary Cryan, Heng Guo, and Giorgos Mousa. Modified log-sobolev inequalities for strongly log-concave distributions. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1358–1370. IEEE, 2019.

Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Anthimos Vardis Kandiros. Learning ising models from one or multiple samples. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 161–168, 2021.

Constantinos Daskalakis and Qinxuan Pan. Sample-optimal and efficient learning of tree ising models. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 133–146, 2021.

Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Optimal phylogenetic reconstruction. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 159–168, 2006.

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

Constantinos Daskalakis, Nishanth Dikkala, and Ioannis Panageas. Logistic regression with peer-group effects via inference in higher-order ising models. In *International Conference on Artificial Intelligence and Statistics*, pages 3653–3663. PMLR, 2020.

Nabarun Deb and Sumit Mukherjee. Fluctuations in mean-field ising models. *The Annals of Applied Probability*, 33(3):1961–2003, 2023.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The minimax learning rates of normal and ising undirected graphical models. 2020.

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

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE, 2022.

Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: fast mixing in high-temperature ising models. *Probability theory and related fields*, 182(3):1035–1051, 2022.

Richard S Ellis. *Entropy, large deviations, and statistical mechanics*. Springer, 2007.

Glenn Ellison. Learning, local interaction, and coordination. *Econometrica: Journal of the Econometric Society*, pages 1047–1071, 1993.

Joseph Felsenstein. *Inferring phylogenies*. Sunderland; Sinauer Associates, 2004.

Joel Friedman. A proof of alon's second eigenvalue conjecture. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 720–724, 2003.

Jason Gaitonde and Elchanan Mossel. A unified approach to learning ising models: Beyond independence and bounded width. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 503–514, 2024.

Jason Gaitonde, Ankur Moitra, and Elchanan Mossel. Bypassing the noisy parity barrier: Learning higher-order markov random fields from dynamics. *arXiv preprint arXiv:2409.05284*, 2024.

Jason Gaitonde, Ankur Moitra, and Elchanan Mossel. Better models and algorithms for learning ising models from dynamics. *arXiv preprint arXiv:2507.15173*, 2025.

Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. Inapproximability for antiferromagnetic spin systems in the tree nonuniqueness region. *Journal of the ACM (JACM)*, 62(6):1–60, 2015.

Andreas Galanis, Alkis Kalavasis, and Anthimos Vardis Kandiros. Learning hard-constrained models with one sample. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3184–3196. SIAM, 2024a.

Andreas Galanis, Alkis Kalavasis, and Anthimos Vardis Kandiros. On sampling from ising models with spectral constraints. *arXiv preprint arXiv:2407.07645*, 2024b.

Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the international congress of mathematicians*, volume 1, page 2. Berkeley, CA, 1986.

Promit Ghosal and Sumit Mukherjee. Joint estimation of parameters in ising model. 2020.

Friedrich Götze, Holger Sambale, and Arthur Sinulis. Concentration inequalities for bounded functionals via log-sobolev-type inequalities. *Journal of Theoretical Probability*, 34:1623–1652, 2021.

Linus Hamilton, Frederic Koehler, and Ankur Moitra. Information theoretic properties of markov random fields, and their algorithmic applications. *Advances in Neural Information Processing Systems*, 30, 2017.

Abhijith Jayakumar, Andrey Y Lokhov, Sidhant Misra, and Marc Vuffray. Discrete distributions are learnable from metastable samples. *arXiv preprint arXiv:2410.13800*, 2024.

Haotian Jiang, Tarun Kathuria, Yin Tat Lee, Swati Padmanabhan, and Zhao Song. A faster interior point method for semidefinite programming. In *2020 IEEE 61st annual symposium on foundations of computer science (FOCS)*, pages 910–918. IEEE, 2020.

Vardis Kandiros, Yuval Dagan, Nishanth Dikkala, Surbhi Goel, and Constantinos Daskalakis. Statistical estimation from dependent data. In *International Conference on Machine Learning*, pages 5269–5278. PMLR, 2021.

Vardis Kandiros, Constantinos Daskalakis, Yuval Dagan, and Davin Choo. Learning and testing latent-tree ising models efficiently. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1666–1729. PMLR, 2023.

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

Frederic Koehler and Thuy-Duong Vuong. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. *arXiv preprint arXiv:2310.01762*, 2023.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. *arXiv preprint arXiv:2210.00726*, 2022a.

Frederic Koehler, Holden Lee, and Andrej Risteski. Sampling approximately low-rank ising models: Mcmc meets variational methods. In *Conference on Learning Theory*, pages 4945–4988. PMLR, 2022b.

Frederic Koehler, Holden Lee, and Thuy-Duong Vuong. Efficiently learning and sampling multimodal distributions with data-based initialization. *arXiv preprint arXiv:2411.09117*, 2024.

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

Dmitriy Kunisky. Optimality of glauber dynamics for general-purpose ising model sampling and free energy approximation. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5013–5028. SIAM, 2024.

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

Holden Lee. Parallelising glauber dynamics. *arXiv preprint arXiv:2307.07131*, 2023.

Yin Tat Lee, Aaron Sidford, and Santosh S Vempala. Efficient convex optimization with membership oracles. In *Conference On Learning Theory*, pages 1292–1294. PMLR, 2018.

Wilhelm Lenz. Beitršge zum verstšndnis der magnetischen eigenschaften in festen kšrpern. *Physikalische Z*, 21(613-615):1, 1920.

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

Kuikui Liu, Sidhanth Mohanty, Amit Rajaraman, and David X Wu. Fast mixing in sparse random ising models. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 120–128. IEEE, 2024.

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

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

Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

Elchanan Mossel, Dror Weitz, and Nicholas Wormald. On the hardness of sampling independent sets beyond the tree threshold. *Probability Theory and Related Fields*, 143(3):401–439, 2009.

Somabha Mukherjee, Ziang Niu, Sagnik Halder, Bhaswar B Bhattacharya, and George Michailidis. High dimensional logistic regression under network dependence. *arXiv preprint arXiv:2110.03200*, 2021.

Somabha Mukherjee, Jaesung Son, and Bhaswar B Bhattacharya. Estimation in tensor ising models. *Information and Inference: A Journal of the IMA*, 11(4):1457–1500, 2022.

Konstantinos E Nikolakakis, Dionysios S Kalogerias, and Anand D Sarwate. Predictive learning on hidden tree-structured ising models. *Journal of Machine Learning Research*, 22(59):1–82, 2021.

Dmitry Panchenko. The sherrington-kirkpatrick model: an overview. *Journal of Statistical Physics*, 149(2):362–383, 2012.

Holger Sambale and Arthur Sinulis. Modified log-sobolev inequalities and two-level concentration. *arXiv preprint arXiv:1905.06137*, 2019.

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

Allan Sly. Computational transition at the uniqueness threshold. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 287–296. IEEE, 2010.

Allan Sly and Nike Sun. The computational hardness of counting in two-spin models on d-regular graphs. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 361–369. IEEE, 2012.

Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2(3):2–3, 2014.

Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. *Advances in neural information processing systems*, 29, 2016.

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

Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. *Advances in Neural Information Processing Systems*, 32, 2019.

Yuanzhe Xu and Sumit Mukherjee. Inference in ising models on dense regular graphs. *The Annals of Statistics*, 51(3):1183–1206, 2023.

Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, and Steven Wu. Privately learning markov random fields. In *International conference on machine learning*, pages 11129–11140. PMLR, 2020.

Appendix A. Preliminaries

In this Section, we collect useful definitions, notations, and facts. Throughout the proofs, we will use generic constant notations such as C, C', K, K'. These refer to absolute constants that could be different from place to place, unless it is specified that some constant depends on some other quantity. For some matrix $A \in \mathcal{S}_0^n$, we denote by $A^{(l)} \in \mathbb{R}^{ln \times ln}$ the corresponding block diagonal matrix with each block equal to A and a total of l blocks. Similarly, for some set of matrices $\mathcal{R} \subseteq \mathbb{R}^{n \times n}$, we denote $\mathcal{R}^{(l)} := \{J^{(l)} : J \in \mathcal{R}\}$. For a symmetric matrix $J \in \mathcal{S}_0^n$ and a subset $I \subseteq [n]$, we denote by $J_I \in \mathbb{R}^{|I| \times n}$ the matrix consisting only of the rows of J that are indexed by elements in I. We also denote $J_{II} \in \mathbb{R}^{|I| \times |I|}$ the submatrix with rows and columns indexed by I.

We start by reviewing properties of the negative log-pseudolikelihood function ϕ, which is defined as

$$\phi(J) := -\log PL(J; X^{(1)}, \ldots, X^{(l)}) = \sum_{k=1}^{l} \sum_{i=1}^{n} \left(\log \cosh(J_i X^{(k)}) - X_i^{(k)} J_i X^{(k)} + \log 2 \right) \quad (6)$$

We define the first and second derivatives of ϕ at a matrix $J \in \mathcal{S}_0^n$ in the direction of a matrix $A \in \mathcal{S}_0^n$ as in Dagan et al. (2021)

$$\left. \frac{\mathrm{d}\phi(J + tA)}{\mathrm{d}t} \right|_{t=0} = \frac{1}{2} \sum_{i=1}^{n} (A_i x)(\tanh(J_i x + h_i) - x_i) \quad (7)$$

$$\left. \frac{\mathrm{d}^2\phi(J + tA)}{\mathrm{d}^2 t} \right|_{t=0} = \frac{1}{2} \sum_{i=1}^{n} (A_i X + h_i)^2 \operatorname{sech}(J_i X + h_i)^2 \quad (8)$$

Since the second derivative is always non-negative, we conclude that ϕ is a convex function. We have included a known external field $h \in \mathbb{R}^n$ in the expression, since the general result does not require 0 external field.

We now introduce some important functional inequalities that will be used in this work. The definitions below are standard and can be found, e.g., in Van Handel (2014). For a measure μ in the hypercube $\{-1, 1\}^n$, we write $\mathbf{E}_\mu, \mathbf{Var}_\mu$ for the expectation and variance with respect to this measure μ. The *Dirichlet form* associated with the Glauber dynamics for a measure μ can be defined as the following operator on two arbitrary functions $f, g : \{-1, 1\}^n \mapsto \mathbb{R}$

$$\mathcal{E}(f, g) := \frac{1}{n} \mathop{\mathbf{E}}_{\mu} \left[\sum_{i=1}^{n} \left(\mathop{\mathbf{E}}_{\mu}[f(X)|X_{-i}] - f(X) \right) \cdot \left(\mathop{\mathbf{E}}_{\mu}[g(X)|X_{-i}] - g(X) \right) \right] \quad (9)$$

We define the *entropy* of a positive function $f : \{-1, 1\}^n \mapsto \mathbb{R}^+$ as

$$\mathrm{Ent}_\mu(f) := \mathop{\mathbf{E}}_{\mu}[f(X) \log f(X)] - \mathop{\mathbf{E}}_{\mu}[f(X)] \cdot \log \mathop{\mathbf{E}}_{\mu}[f(X)] \quad (10)$$

We will need two notions of a modified log-Sobolev inequality. To define the first one, let $\mathcal{G}$ be the set of all functions $\{-1, 1\}^n \mapsto \mathbb{R}$ and $\mathcal{G}'$ the subset of all positive functions. Let $\Gamma : \mathcal{G} \mapsto \mathcal{G}'$ be a so-called *difference operator*. We say a measure μ supported on the hypercube satisfies a Γ-MLSI(ρ), for some $\rho > 0$, if and only if for all $f \in \mathcal{G}$

$$\mathrm{Ent}_\mu\left(e^f\right) \leq \frac{\rho}{2} \mathop{\mathbf{E}}_{\mu}\left[\Gamma(f)^2 e^f\right] \quad ,$$

This notation is a classical notion of the Modified Log-Sobolev inequality and has been used in prior work for proving concentration inequalities Bobkov and Götze (1999); Sambale and Sinulis (2019).

There is a related notion of MLSI based on the Glauber dynamics. We say μ satisfies the *modified log-Sobolev inequality for the Glauber dynamics* with constant $C > 0$ and write Glauber-MLSI(C), if for every function $f : \{-1, 1\}^n \mapsto \mathbb{R}$

$$\mathrm{Ent}_\mu\left(e^f\right) \leq C \cdot n \cdot \mathcal{E}(e^f, f)$$

This version of the Modified Log-Sobolev inequality arises naturally when considering the rate of contraction of the KL-divergence between the distribution of a Markov Chain at a given time and the stationary distribution of the chain Bobkov and Tetali (2006).

Also, we say a measure μ satisfies a *Poincare inequality* with constant $\rho > 0$ and write Po(ρ), if for every function $f : \{-1, 1\}^n \mapsto \mathbb{R}$

$$\operatorname*{\mathbf{Var}}_\mu(f) \leq \rho \cdot n \cdot \mathcal{E}(f, f) \tag{11}$$

Both the Poincaré inequality and Glauber-MLSI have been established in prior work for spectrally bounded Ising models.

Lemma 11 (Eldan et al. (2022)) *If $\lambda_{\max}(J^*) - \lambda_{\min}(J^*) < 1 - \alpha$, then $\mathbf{Pr}_{J^*}$ satisfies Po$(1/\alpha)$.*

Lemma 12 (Anari et al. (2021); Chen and Eldan (2022)) *If $\lambda_{\max}(J^*) - \lambda_{\min}(J^*) < 1 - \alpha$, then $\mathbf{Pr}_{J^*}$ satisfies Glauber-MLSI$(1/\alpha)$.*

Furthermore, it is a well-known identity (for example, Bobkov and Tetali (2006)) that, in general, the Glauber MLSI implies the Poincaré inequality with a slightly worse constant. For completeness, we give a proof below.

Lemma 13 *If for some μ, Glauber-MLSI(C) satisfies, then it also satisfies Po$(2C)$.*

Proof Consider $F(t) = Cn\mathcal{E}(e^{tf}, tf) - \mathrm{Ent}_\mu\left(e^{tf}\right)$. We know that $F(0) = 0$. Since we have $F \geq 0$, if we have $F'(0) = 0$, we must have $F''(0) \geq 0$. Taking the derivative, we have

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}t^2} n\mathcal{E}(e^{tf}, tf) &= \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(C \mathop{\mathbf{E}}_\mu\left[\sum_{i=1}^n \left(\mathop{\mathbf{E}}_\mu[e^{tf}|X_{-i}] - e^{tf}\right)\left(\mathop{\mathbf{E}}_\mu[tf|X_{-i}] - tf\right)\right]\right) \\
&= C\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathop{\mathbf{E}}_\mu\left[\sum_{i=1}^n \left(\mathop{\mathbf{E}}_\mu[fe^{tf}|X_{-i}] - fe^{tf}\right)\left(\mathop{\mathbf{E}}_\mu[tf|X_{-i}] - tf\right)\right.\right. \\
&\qquad \left.\left. + \left(\mathop{\mathbf{E}}_\mu[e^{tf}|X_{-i}] - e^{tf}\right)\left(\mathop{\mathbf{E}}_\mu[f|X_{-i}] - f\right)\right]\right) \\
&= C\left(\mathop{\mathbf{E}}_\mu\left[\sum_{i=1}^n \left(\mathop{\mathbf{E}}_\mu[f^2e^{tf}|X_{-i}] - f^2e^{tf}\right)\left(\mathop{\mathbf{E}}_\mu[tf|X_{-i}] - tf\right)\right.\right. \\
&\qquad \left.\left. + 2\left(\mathop{\mathbf{E}}_\mu[fe^{tf}|X_{-i}] - fe^{tf}\right)\left(\mathop{\mathbf{E}}_\mu[f|X_{-i}] - f\right)\right]\right)
\end{aligned}$$

When we set $t = 0$, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{E}(e^{tf}, tf)\Big|_{t=0} = 2\mathop{\mathbf{E}}_{\mu}\left(\left[\left(\mathop{\mathbf{E}}_{\mu}[f|X_{-i}] - f\right)\left(\mathop{\mathbf{E}}_{\mu}[f|X_{-i}] - f\right)\right]\right) = 2n\mathcal{E}(f, f)$$

Also, we have

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathrm{Ent}_{\mu}\left(e^{tf}\right) &= \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\mathop{\mathbf{E}}_{\mu}[e^{tf}\cdot tf] - \mathop{\mathbf{E}}_{\mu}[e^{tf}]\log\mathop{\mathbf{E}}_{\mu}[e^{tf}]\right)\\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathop{\mathbf{E}}_{\mu}[e^{tf}\cdot f + e^{tf}\cdot tf^2] - \mathop{\mathbf{E}}_{\mu}[fe^{tf}]\log\mathop{\mathbf{E}}_{\mu}[e^{tf}] - \mathop{\mathbf{E}}_{\mu}[fe^{tf}]\right)\\ &= \left(\mathop{\mathbf{E}}_{\mu}[2e^{tf}\cdot f^2 + e^{tf}\cdot tf^3] - \mathop{\mathbf{E}}_{\mu}[f^2e^{tf}]\log\mathop{\mathbf{E}}_{\mu}[e^{tf}] - \frac{\mathbf{E}_{\mu}[fe^{tf}]^2}{\mathbf{E}_{\mu}[e^{tf}]^2} - \mathop{\mathbf{E}}_{\mu}[f^2e^{tf}]\right)\end{aligned}$$

Therefore, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathrm{Ent}_{\mu}\left(e^{tf}\right)\Big|_{t=0} = \left(2\mathop{\mathbf{E}}_{\mu}[f^2] - \mathop{\mathbf{E}}_{\mu}[f^2]^2 - \mathop{\mathbf{E}}_{\mu}[f]^2\right) = \mathop{\mathbf{Var}}_{\mu}(f)$$

Therefore, we have $F''(0) = 2Cn\mathcal{E}(f, f) - \mathbf{Var}_{\mu}(f) \geq 0$, which means that Po$(2C)$ holds. ■

The reason Γ-MLSI is useful is that concentration results for second-order polynomials are proven in Sambale and Sinulis (2019) when the distribution μ satisfies this property. On the other hand, Glauber-MLSI is usually the property that follows from results about fast mixing Eldan et al. (2022); Chen and Eldan (2022); Anari et al. (2021). It would thus be desirable to connect the two notions of MLSI in order to establish concentration results for the Ising model. To facilitate that connection, following Sambale and Sinulis (2019) we define the operator $\mathfrak{d}^+$ as follows

$$\mathfrak{d}^+ f(X) = \sqrt{\sum_{i=1}^{n}\mathbf{E}\left[((f(X) - f(X_i', X_{-i}))_+^2\Big|X_{-i}\right]}\ . \tag{12}$$

In the above, we use (X_i', X_{-i}) be a shortened random vector $(X_1', X_2', \ldots, X_n')$ that has all other coordinate j, $X_j' = X_j$, and X_i' is sampled independently of everything else according to the distribution conditioning on $X_{-i}' = X_{-i}$. Also, we have used the shorthand notation $x_+ := \max(0, x)$. The quantity $\mathfrak{d}^+ f(x)$ can be thought of as the l_2-norm of a discrete derivative of f at x. It is therefore capturing the Lipschitzness of f in some appropriate sense and will thus be important for proving that $f(X)$ concentrates. The connection between the two definitions is that Glauber-MLSI(ρ) implies $\mathfrak{d}^+$-MLSI(2ρ), since by direct calculation we can get that $\mathbf{E}_{\mu}[\mathfrak{d}^+ f \cdot e^f] \geq n\mathcal{E}(f, e^f)$ Sambale and Sinulis (2019). Indeed, in the following lemma, we use this fact and derive a generic concentration result involving $\mathfrak{d}^+$.

Lemma 14 *Suppose that μ satisfy Glauber-MLSI(ρ). Then, for any functions $f, g : \{-1, 1\}^n \mapsto \mathbb{R}$ and constant $b > 0$ such that $\mathfrak{d}^+ f \leq g$ and $\mathfrak{d}^+ g \leq b$, we have for any $t > 0$*

$$\mathbf{Pr}\left[|f(X) - \mathbf{E}[f(X)]| > t\right] \leq \frac{8}{3}\exp\left(-\frac{1}{16\rho}\min\left(\frac{t^2}{\mathbf{E}[g]^2}, \frac{t}{b}\right)\right).$$

Proof By Proposition 2.18 in Sambale and Sinulis (2019), if μ satisfies Glauber-MLSI$(1/\alpha)$, it follows that the same Ising model also satisfies $\mathfrak{d}^+-$MLSI$(2/\alpha)$. Thus, we can apply Corollary 1.2 from Sambale and Sinulis (2019) and the result immediately follows. ■

We now state a useful lemma for upper-bounding second moments of the model.

Lemma 15 *Let A be any matrix, and $x \sim \mu$ be an Ising model satisfying Po(ρ). We have the following:*

$$\mathbf{E}\Big(\|Ax\|^2\Big) \leq \|\,\mathbf{E}[Ax]\|^2 + \rho\|A\|_F^2$$

Proof By the Poincaré Inequality, we have that

$$\mathbf{E}[(A_i x)^2] - (\mathbf{E}[A_i x])^2 = \mathbf{Var}(A_i x) \leq \rho \mathop{\mathbf{E}}_{\mu} \sum_{j=1}^{n} (\mathbf{E}[A_i x | x_{-j}] - A_i x)^2 \leq \rho \sum_{j=1}^{n} A_{ij}^2.$$

Therefore, we have in total

$$\mathbf{E}(\|Ax\|^2) \leq \sum_{i=1}^{n} (\mathbf{E}[A_i x])^2 + \rho \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}^2 = \|\,\mathbf{E}[Ax]\|^2 + \rho\|A\|_F^2\,.$$

■

Another useful consequence of MLSI is the simpler concentration of Lipschitz functions, which is well known. Below is one version of this implication from Cryan et al. (2019).

Lemma 16 (Lemma 15 in Cryan et al. (2019)) *Let P be the transition matrix of a reversible Markov Chain with stationary distribution π on a finite set Ω, and $f : \Omega \to \mathbb{R}$ be some observable function. Then,*

$$\mathop{\mathbf{Pr}}_{x\sim\pi}(f(x) - \mathop{\mathbf{E}}_{\pi} f \geq a) \leq \exp(-\frac{a^2}{2n \cdot \rho \cdot v(f)})$$

where $a \geq 0$ and

$$v(f) := \max_{x\in\Omega} \left\{ \sum_{y\in\Omega} P(x,y)(f(x) - f(y))^2 \right\} \quad ,$$

if π satisfies Glauber-MLSI$(\rho(P))$.

Finally, we also require the *Hubbard-Stratonovich* transform, which is a way to decompose any Ising model distribution into a mixture of product distributions. Formally, suppose $X \sim \mathbf{Pr}_{J^*}$ and $G \sim \mathcal{N}(0, I_n)$. Let us consider the random variable

$$Y = X + J^{-1/2}G \tag{13}$$

An easy calculation now shows that the distribution of X conditioned on Y is an Ising model with *zero* interaction matrix and external field J^*Y (for details see e.g. Theorem 3.12 in Liu et al. (2024)). Thus, if π_y is the distribution of Y, we can decompose any Ising measure as

$$\mathop{\mathbf{Pr}}_{J^*}[x] = \int_{y\in\mathbb{R}^n} \pi(y) \mathop{\mathbf{Pr}}_{0,J^*y}[x]dy \tag{14}$$

Appendix B. Learning Ising Models with Bounded Operator Norm

In this Section, we study Ising models of the form (1) with a known and bounded external field h whose interaction matrix has bounded operator norm and satisfy the modified log-Sobolev Inequality. In particular, throughout the section, we will make the following assumption.

Assumption 17 *The set of all candidate matrices, $\mathcal{R} \subseteq \mathcal{S}_0^n$, contains J^*. The Ising model, as in (1) with h a known external field, satisfies $\|J^*\|_{op} \leq \lambda$ for some constant $\lambda > 0$. Also, it satisfies Glauber-MLSI(ρ). Finally, $\max_{1\leq i\leq n} |h_i| \leq h_{\max}$.*

Without loss of generality, we assume $\rho, \lambda \geq 1$.

B.1. Concentration of the first derivative

The first derivative of the pseudolikelihood is given by the formula in (7). Even though this function is not a polynomial, intuitively it behaves similarly to a second-degree polynomial if we linearize the $\tanh$ function. Indeed, our goal in this section will be to prove that it concentrates similarly to a second-degree polynomial. This was shown to hold in Dagan et al. (2021) in the case where $\|J^*\|_\infty < 1$, using the Approximate Tensorization of Entropy (ATE) and the bound on the infinity norm. Here, we will show that it still holds under MLSI and bounded operator norm, using Lemma 14.

Theorem 18 *Suppose $X \in \{-1,1\}^n$ is sampled from an Ising model with interaction matrix J^* satisfying Assumption 17. For a fixed vector $b \in \mathbb{R}^n$ and fixed symmetric matrix $A \in \mathbb{R}^{n\times n}$ with zero diagonal, let us define the function*

$$f(X) = \sum_{i=1}^{n} (A_i X + b_i)(\tanh(J_i^* X + h_i) - X_i) \ .$$

Then, we can take $C = \frac{1}{2^{21}\lambda^4\rho^2} > 0$, such that for all $t > 0$

$$\mathbf{Pr}\left[|f(X)| > t\right] \leq \frac{8}{3}\exp\left(-C\min\left(\frac{t^2}{\|A\|_F^2 + \|\mathbf{E}[AX+b]\|_2^2}, \frac{t}{\|A\|_{op}}\right)\right) \ .$$

Proof Let us define for a vector $X \in \{-1,1\}^n$ the vector $\overline{X^{(i)}} \in \{-1,1\}^n$, where the i-th coordinate is flipped. Also, define vectors $X^{(k+)}, X^{(k-)}$ which $X_j^{(k+)} = X_j^{(k-)} = X_j$ holds for all coordinates $j = 1,\ldots,n, j \neq k$, while $X_k^{(k+)} = 1, X_k^{(k-)} = -1$. Moreover, let us define the matrix $W = W(X) \in \mathbb{R}^{n\times n}$, where the element in the jth row, kth column is equal to

$$W_{jk} = \tanh(J_j^* X^{(k+)} + h) - \tanh(J_j^* X^{(k-)} + h) - 2J_{jk}^*,$$

In the sequel, we might omit the dependence of W on X for brevity. Note that W is not necessarily symmetric. We denote by W_k the kth column and W^k the k-th row of W. First, Let us bound $\|W\|_{op}$, where $\|W\|_{op} = \sqrt{\|WW^\top\|_{op}}$ Note the well-knownn fact that for all $a, b \in \mathbb{R}$

$$|\tanh(a) - \tanh(b) - (a-b)| \leq \frac{1}{2}(a-b)^2 \tag{15}$$

Then using (15), we can bound each entry W_{jk} of the matrix as

$$|W_{jk}| \leq \frac{1}{2}(2J^*_{jk})^2 = 2(J^*_{jk})^2 \ ,$$

Thus, the l_1 norm of every column j of $W(X)$ can be bounded as follows

$$\|W_j\|_1 = \sum_{k=1}^{n} W_{jk} \leq 2\sum_{k=1}^{n}(J^*_{jk})^2 = 2\|J^*_j\|_2^2 \leq 2\lambda^2 \ .$$

The last inequality follows from the fact that the l_2 norm of every row of J^* is bounded by $\|J^*\|_{op}$. Thus $\|W\|_\infty \leq 2\lambda^2$. Similarly, we can get that $\|W\|_1 \leq 2\lambda^2$, which implies that

$$\|W\|_{op} = \sqrt{\|WW^\top\|_{op}} \leq \sqrt{\|WW^\top\|_1} \leq \sqrt{\|W\|_\infty \cdot \|W\|_1} \leq 2\lambda^2 \ .$$

We start by bounding $\mathfrak{d}^+ f$. We first notice that for any given X (again, $\mathbf{E}$ in the following expression is same as Equation (12): X'_i is sampled according to the conditional distribution with fixed X_{-i}),

$$\begin{aligned}
(\mathfrak{d}^+ f(X))^2 &= \sum_{i=1}^{n} \mathbf{E}\left[\left((f(X) - f(X'_i, X_{-i}))_+^2 \middle| X_{-i}\right]\right. \\
&= \sum_{i=1}^{n} \mathbf{Pr}[X'_i = 1 - X_i | X_{-i}](f(X) - f(X_{-i}, 1 - X_i))_+^2 \\
&\leq \sum_{i=1}^{n} \left(f(X) - f(\overline{X^{(i)}})\right)^2
\end{aligned}$$

We also define the vector functions

$$g(x) := Ax + b \quad , \quad h(x) := \tanh(J^* x + b) - x \ ,$$

which allows us to write $f(x) = g(x)^\top h(x)$. Here, we write $h(x)$ as a vector, with $h(x)_i = \tanh(J^*_i x + b) - x_i$. Therefore, we can calculate that

$$h(x)_j - h(\overline{x^{(i)}})_j = \tanh(J^*_j x + b) - \tanh(J^*_j \overline{x^{(i)}} + b) - 2x_i \mathbb{1}(i = j).$$

So, we have

$$h(x) - h(\overline{x^{(i)}}) = \tanh(J^* x + b) - \tanh(J^* \overline{x^{(i)}} + b) - 2x_i e_i = x_i W_i + 2x_i J^*_i - 2x_i e_i = x_i(W_i + 2J^*_i - 2e_i),$$

where e_i is the standard basis vector with all coordinates 0 except for the i-th that is 1. With that, we can further bound the discrete derivative as follows. For all $x \in \{-1, 1\}^n$, we have

$$\begin{aligned}
&(\mathfrak{d}^+ f(x))^2 \\
\leq &\sum_{i=1}^{n} \left(g(x)^\top h(x) - g(\overline{x^{(i)}})^\top h(\overline{x^{(i)}})\right)^2 \\
= &\sum_{i=1}^{n} \left(\left(g(x) - g(\overline{x^{(i)}})\right)^\top h(x) + \left(g(\overline{x^{(i)}}) - g(x)\right)^\top \left(h(x) - h(\overline{x^{(i)}})\right) + g(x)^\top \left(h(x) - h(\overline{x^{(i)}})\right)\right)^2 \\
= &\sum_{i=1}^{n} \left(2x_i A_i^\top h(x) + 2x_i A_i^\top \cdot x_i \left(2J^*_i + W_i - 2e_i\right) + (Ax + b)^\top \cdot x_i \left(2J^*_i + W_i - 2e_i\right)\right)^2 .
\end{aligned}$$

By Cauchy-Schwartz inequality, $(x+y+z)^2 \le 3(x^2+y^2+z^2)$, and we know that $x_i = \pm 1$, we have the above expression is no larger than

$$3\sum_{i=1}^{n}\left(4\left(A_i^\top h(x)\right)^2 + 4\left(A_i^\top(2J_i^* + W_i - 2e_i)\right)^2 + \left((Ax+b)^\top(2J_i^* + W_i - 2e_i)\right)^2\right)$$

First, we deal with the middle term $\left(A_i^\top(2J_i^* + W_i - 2e_i)\right)^2$. Using Cauchy-Schwartz inequality for vectors ($x^\top y \le \|x\| \cdot \|y\|$), for the middle term, we have

$$\begin{aligned}&\left(A_i^\top(2J_i^* + W_i - 2e_i)\right)^2 \le \|A_i\|_2^2 \cdot \|2J_i^* + W_i - 2e_i\|_2^2\\ \le&\|A_i\|_2^2 \cdot (2\|J_i^*\|_2 + \|W_i\|_2 + 2)^2 \le (2+2\lambda+2\lambda^2)^2\|A_i\|^2.\end{aligned}$$

Here, the last inequality is because $\|J_i\|_2 \le \lambda$ and $\|W_i\|_2 \le 2\lambda^2$. Therefore, summing over i, we have

$$3\sum_{i=1}^{n} 4\left(A_i^\top(2J_i^* + W_i - 2e_i)\right)^2 \le 3\sum_{i=1}^{n} 4\cdot(2+2\lambda+2\lambda^2)^2\|A_i\|^2 = 48(1+\lambda+\lambda^2)^2\|A\|_F^2.$$

Now, we deal with the last term. $\left((Ax+b)^\top(2J_i^* + W_i - 2e_i)\right)^2$ By Cauchy-Schwartz inequality again, for the last term, we have

$$\begin{aligned}&\left((Ax+b)^\top(2J_i^* + W_i - 2e_i)\right)^2\\ \le&12\left(A_i^\top x + b_i\right)^2 + 12\left((Ax+b)^\top J_i^*\right)^2 + 3\left((Ax+b)^\top W_i\right)^2\end{aligned}$$

Again, after summing over i, then using the fact that $\|J^*\|_{op} \le \lambda$ and $\|W\|_{op} \le 2\lambda^2$, we can deduce that

$$\begin{aligned}&\sum_{i=1}^{n} 12\left(A_i^\top x + b_i\right)^2 + 12\left((Ax+b)^\top J_i^*\right)^2 + 3\left((Ax+b)^\top W_i\right)^2\\ =&12\|(Ax+b)^\top J_i^*\|^2 + 3\|(Ax+b)^\top W\|^2 + 12\|Ax+b\|_2^2\\ \le&12\lambda^2\|Ax+b\|_2^2 + 12\lambda^4\|Ax+b\|_2^2 + 12\|Ax+b\|_2^2\\ =&12(1+\lambda^2+\lambda^4)\|Ax+b\|^2 = 12(1+\lambda^2+\lambda^4)\sum_{i=1}^{n}(A_i^\top x + b_i)^2\end{aligned}$$

Combining them together, we have

$$\begin{aligned}(\mathfrak{d}^+ f(x))^2 \le&3\sum_{i=1}^{n}\left(4\left(A_i^\top h(x)\right)^2 + 4\left(A_i^\top(2J_i^* + W_i - 2e_i)\right)^2 + \left((Ax+b)^\top(2J_i^* + W_i - 2e_i)\right)^2\right)\\ \le&48(1+\lambda+\lambda^2)^2\|A\|_F^2 + (3\cdot 12)(1+\lambda^2+\lambda^4)\sum_{i=1}^{n}(A_i^\top x + b_i)^2 + 3\sum_{i=1}^{n} 4\left(A_i^\top h(x)\right)^2\\ \le&48(1+\lambda+\lambda^2)^2\|A\|_F^2 + 36(1+\lambda^2+\lambda^4)\sum_{i=1}^{n}(A_i^\top x + b_i)^2 + \left(A_i^\top h(x)\right)^2\end{aligned}$$

Let

$$q(x) = \sqrt{48(1+\lambda+\lambda^2)^2\|A\|_F^2 + 36(1+\lambda^2+\lambda^4)\sum_{i=1}^n (A_i^\top x + b_i)^2 + \left(A_i^\top h(x)\right)^2}. \tag{16}$$

We thus have $\mathfrak{d}^+ f(x) \le q(x)$ for all $x \in \{-1,1\}^n$. Let us first bound $\mathbf{E}\left[q(X)\right]^2 \le \mathbf{E}[q(x)^2]$. First, notice that

$$\mathbf{E}[h(X)] = \mathbf{E}[X - \tanh(J^* X + h)] = 0 \ .$$

This enables us to write.

$$\mathbf{E}[(A_i^\top h(X))^2] = \mathbf{Var}[A_i^\top h(X)]$$

We can now use the Poincaré inequality to bound the above variance. Note that by Lemma 13, if μ satisfies Glauber MLSI(ρ), then it also satisfies Po(2ρ). So we get

$$\begin{aligned}
\mathbf{Var}[A_i^\top h(X)] &\le 2\rho \sum_{k=1}^n \mathbf{E}\left[\mathbf{Var}\left[A_i^\top h(X)|X_{-k}\right]\right] \\
&\le 2\rho \sum_{k=1}^n \mathbf{E}\left[\left(A_i^\top (h(X_{k+}) - h(X_{k-}))\right)^2\right]
\end{aligned}$$

We have that

$$\begin{aligned}
A_i^\top (h(X_{k+}) - h(X_{k-})) &= A_i^\top (\tanh(J^* X_{k+} + h) - \tanh(J^* X_{k-} + h) - 2e_k) \\
&= -2A_{ik} + A_i^\top (2J^* e_k) + A_i^\top W_k \\
&= -2A_{ik} + 2A_i^\top J_k^* + A_i^\top W_k
\end{aligned}$$

Applying the Cauchy-Schwarz inequality now yields

$$\left(A_i^\top (h(X_{k+}) - h(X_{k-}))\right)^2 \le 3\left(4A_{ik}^2 + 4(A_i^\top J_k^*)^2 + (A_i^\top W^k)^2\right)$$

Summing over all k now gives

$$\begin{aligned}
\sum_{k=1}^n \left(A_i^\top (h(X_{k+}) - h(X_{k-}))\right)^2 &\le 3\sum_{k=1}^n \left(4A_{ik}^2 + 4(A_i^\top J_k^*)^2 + (A_i^\top W^k)^2\right) \\
&= 12\|A_i\|_2^2 + 12\|J^* A_i\|_2^2 + 3\|W^\top A_i\|_2^2 \\
&\le 12\|A_i\|_2^2 + 12\|J^{*\top}\|_{op}^2 \|A_i\|_2^2 + 3\|W^\top\|_{op}^2 \|A_i\|_2^2.
\end{aligned}$$

Using the fact that $\|J^*\|_{op} < \lambda$ and $\|W\|_{op} \le 2\lambda^2$, we have

$$\sum_{k=1}^n \left(A_i^\top (h(X_{k+}) - h(X_{k-}))\right)^2 \le 12(1+\lambda^2+\lambda^4)\|A_i\|_2^2.$$

Therefore, we have

$$\mathbf{E}[(A_i^\top h(X))^2] \le 24\rho(1+\lambda^2+\lambda^4)\|A_i\|^2 \ .$$

For the second term, the analysis follows along similar lines if we apply a suitable centering to make the variance appear. In particular, we can write

$$\begin{aligned}\mathbf{E}[(A_i^\top X + b_i)^2] - \mathbf{E}\left[A_i^\top X + b_i\right]^2 &= \mathbf{Var}[A_i^\top X] \\ &\le 2\rho \sum_{k=1}^n \mathbf{E}\left[\mathbf{Var}[A_i^\top X | X_{-i}]\right] \\ &\le 2\rho \sum_{k=1}^n A_{ik}^2 \\ &= 2\rho \|A_i\|_2^2\end{aligned}$$

Summing over all i and using (16) gives

$$\begin{aligned}\mathbf{E}[q(X)]^2 &\le 48(1+\lambda+\lambda^2)^2\|A\|_F^2 + 36(1+\lambda^2+\lambda^4)\sum_{i=1}^n (A_i^\top x + b_i)^2 + \left(A_i^\top h(x)\right)^2 \\ &\le 48(1+\lambda+\lambda^2)^2\|A\|_F^2 \\ &+ 36(1+\lambda^2+\lambda^4)\sum_{i=1}^n 24\rho(1+\lambda^2+\lambda^4)\|A_i\|^2 + 2\rho(\|A_i\|^2 + \mathbf{E}[A_i^\top X + b]^2) \\ &\le 2^{11}(1+\lambda^4)(1+\rho)\left(\|A\|_F^2 + \|\mathbf{E}[Ax+b]\|_2^2\right). \end{aligned} \tag{17}$$

We will now focus on bounding $\mathfrak{d}^+ q(x)$. We have

$$\begin{aligned}q(x) =& \sqrt{48(1+\lambda+\lambda^2)^2\|A\|_F^2 + 36(1+\lambda^2+\lambda^4)\sum_{i=1}^n (A_i^\top x + b_i)^2 + \left(A_i^\top h(x)\right)^2} \\ \le& \sqrt{48}(1+\lambda+\lambda^2)\|A\|_F + 6\sqrt{1+\lambda^2+\lambda^4}\sqrt{\sum_{i=1}^n (A_i^\top x + b_i)^2 + \left(A_i^\top h(x)\right)^2}.\end{aligned}$$

Let $r(x) = \sqrt{\sum_{i=1}^n (A_i^\top x + b_i)^2 + \left(A_i^\top h(x)\right)^2}$. Then we define

$$K^{(i)}(x) = A_i^\top h(x); \quad L^{(i)}(x) = A_i^\top x + b.$$

We first notice that $\mathfrak{d}^+ q(x) \le 6\sqrt{1+\lambda^2+\lambda^4} \cdot \mathfrak{d}^+ r(x)$. This is because for all positive a, b, c, we have $|\sqrt{a+b} - \sqrt{a+c}| \le |\sqrt{b} - \sqrt{c}|$. We define the vector function $s = s(x) \in \mathbb{R}^{2n}$, with $s(x)_i = K^{(i)}(x)$ if $i \le n$ and $s(x)_i = L^{(i-n)}(x)$ if $i > n$. We then have that

$$r(x) = \|s(x)\|_2 = \sup_{v \in \mathcal{S}^{2n-1}} \langle s(x), v\rangle \ .$$

Let us denote $\tilde{v} = (\tilde{v}_1, \tilde{v}_2) := \operatorname{argmax}_{v \in \mathcal{S}^{2n-1}} \langle s(x), v \rangle$ the unit vector (in $\mathbb{R}^{2n}$, where $\tilde{v}_1, \tilde{v}_2 \in \mathbb{R}^n$) that is in the direction of $s(x)$. Then, we can write

$$\begin{aligned}
(\mathfrak{d}^+ \|r(x)\|)^2 &\leq \sum_{k=1}^{n} \left(\|s(x)\| - \|s(\overline{x^{(k)}})\| \right)_+^2 \\
&= \sum_{k=1}^{n} \left(\langle s(x), \tilde{v} \rangle - \sup_{v \in \mathcal{S}^{2n-1}} \langle s(\overline{x^{(k)}}), v \rangle \right)_+^2 \\
&\leq \sum_{k=1}^{n} \left(\langle s(x), \tilde{v} \rangle - \langle s(\overline{x^{(k)}}), \tilde{v} \rangle \right)_+^2 \\
&\leq \|\boldsymbol{S} \cdot \tilde{v}\|^2 \ ,
\end{aligned}$$

where $\boldsymbol{S} \in \mathbb{R}^{n \times 2n}$ is the matrix with k-th row equal to $s(x) - s(\overline{x^{(k)}})$. We can write in block form $\boldsymbol{S} = (\boldsymbol{S}_1 | \boldsymbol{S}_2)$, where $\boldsymbol{S}_1, \boldsymbol{S}_2 \in \mathbb{R}^{n \times n}$. Let us bound $\|\boldsymbol{S}_1\|_{op}, \|\boldsymbol{S}_2\|_{op}$. We have that

$$\begin{aligned}
(\boldsymbol{S}_1)_{ki} &= s_1^{(i)}(x) - s_1^{(i)}(\overline{x^{(k)}}) \\
&= A_i^\top (h(x) - h(\overline{x^{(k)}})) \\
&= x_k A_i^\top (2J_k^* + W_k - 2e_k) \\
&= x_k (2(J^* A)_{ki} + (WA)_{ki} - 2A_{ki})
\end{aligned}$$

For a vector $x \in \mathbb{R}^n$, we denote $\operatorname{diag}(x)$ the diagonal matrix with diagonal entries equal to x. Then, we can write in matrix form

$$\begin{aligned}
\|\boldsymbol{S}_1\|_{op} &= \|\operatorname{diag}(x)(2J^* A + WA - 2A)\|_{op} \\
&\leq \|\operatorname{diag}(x)\|_{op} \cdot \|2J^* A + WA - 2A\|_{op} \\
&\leq 2\|A\|_{op} + \|J^*\|_{op} \cdot \|A\|_{op} + \|W\|_{op} \cdot \|A\|_{op} \\
&\leq (2 + \lambda + 2\lambda^2)\|A\|_{op} \ .
\end{aligned}$$

For $\boldsymbol{S}_2$ the situation is similar and we can write

$$\begin{aligned}
(\boldsymbol{S}_2)_{ki} &= A_i^\top (x - \overline{x^{(k)}}) \\
&= 2x_k A_{ki} \ .
\end{aligned}$$

Thus

$$\|\boldsymbol{S}_2\|_{op} = 2\|\operatorname{diag}(x) A\|_{op} \leq 2\|A\|_{op}$$

We can now conclude

$$\begin{aligned}
\|\boldsymbol{S} \cdot \tilde{v}\|^2 &= \|\boldsymbol{S}_1 \tilde{v}_1 + \boldsymbol{S}_2 \tilde{v}_2\|^2 \\
&\leq 2\|\boldsymbol{S}_1 \tilde{v}_1\|^2 + 2\|\boldsymbol{S}_2 \tilde{v}_2\|^2 \\
&\leq 2\|\boldsymbol{S}_1\|_{op}^2 \|\tilde{v}_1\|^2 + 2\|\boldsymbol{S}_2\|_{op}^2 \|\tilde{v}_2\|^2 \\
&\leq 2\max(\|\boldsymbol{S}_1\|_{op}^2, \|\boldsymbol{S}_1\|_{op}^2) \leq (2 + \lambda + 2\lambda^2)^2 \|A\|_{op}^2 \ ,
\end{aligned}$$

where we used the fact that $\|\tilde{v}_1\|^2 + \|\tilde{v}_2\|^2 \leq 1$.

To summarize, we have shown that $\mathfrak{d}^+ f(x) \le q(x)$ for all x. Furthermore, by (17) we proved that

$$\mathbf{E}[q]^2 \le 2^{11}(1+\lambda^4)(1+\rho)\left(\|A\|_F^2 + \|\mathbf{E}[Ax+b]\|_2^2\right)\|A\|_F^2 \quad , \quad \mathfrak{d}^+ q(x) \le (2+\lambda+2\lambda^2)^2\|A\|_{op}$$

We also know that

$$\mathbf{E}\left[\frac{\partial\phi(J^*)}{\partial A}\right] = 0$$

Thus, applying Lemma 14, take $C = \frac{1}{16\rho}\frac{1}{2^{15}(1+\lambda^4)(1+\rho)} \ge \frac{1}{2^{21}\lambda^4\rho^2}$ we obtain that for every $t > 0$

$$\mathbf{Pr}\left[\left|\frac{\partial\phi(J^*)}{\partial A}\right| > t\right] \le \frac{8}{3}\exp\left(-C\min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|_{op}}\right)\right) \quad .$$

■

We are now ready to establish the concentration of the first derivative for Ising models satisfying MLSI. This basically follows as a direct Corollary of Theorem 18

Lemma 19 *Suppose $X \in \{-1,1\}^n$ is sampled from an Ising model satisfying Assumption 17. Then, if ϕ is the pseudolikelihood function evaluated at X and $A \in \mathcal{S}_0^n$, then there is a constant $C' = \frac{1}{2^{21}\lambda^4\rho^2}$ such that for any $t > 0$:*

$$\left|\frac{\partial\phi(J^*)}{\partial A}\right| \le t\|A\|_F$$

with probability at least

$$1 - \frac{8}{3}\exp\left(-C'\min\left(t^2, \frac{t\|A\|_F}{\|A\|_{op}}\right)\right)$$

Proof We apply Theorem 18 and substitute $t\|A\|_F$ for t, completing the proof. ■

B.2. Anti-Concentration of the Second Derivative

We want to show that the second derivative is lower bounded with high probability. First, we present the lemma, as the second form of Corollary 1.2 in Sambale and Sinulis (2019) (the inequality right after Corollary 1.2 in Sambale and Sinulis (2019), which does not have a specific number)

Lemma 20 (Corollary 1.2 in Sambale and Sinulis (2019)) *Assume that μ satisfies a Γ-MLSI(ρ) for some difference operator Γ and $\rho > 0$. Let f, g be two measurable functions such that $\Gamma(f) \le g$ and $\Gamma(g) \le b$. Then there is a universal constant c such that for all $t \ge 0$ we have*

$$\mathbf{Pr}[f - \mathop{\mathbf{E}}_{\mu}[f] \ge t] \le \exp\left(-c\min\left(\frac{t^2}{\rho(\mathbf{E}_\mu g)^2 + 2b^2\rho^2}, \frac{t}{\sqrt{2}\rho b}\right)\right). \tag{18}$$

From (8), it is clear that the second derivative is almost equal to a second-degree polynomial, up to a factor that involves the sech function. We thus start by stating a concentration bound for second-degree polynomials that essentially follows from Sambale and Sinulis (2019).

Lemma 21 *Suppose $X \in \{-1,1\}^n$ is sampled from an Ising model with interaction matrix J^* satisfying Glauber-MLSI(ρ) and external field $h \in \mathbb{R}^n$. Also, let $S \in \mathcal{S}_0^n$. Then, there exists an absolute constant $c > 0$, such that for any $t > 0$,*

$$\mathbf{Pr}[x^\top Sx - \mathbf{E}[x^\top Sx] \geq t] \leq \exp\left(-\frac{c}{\rho^2}\min\left(\frac{t^2}{\|S\|_F^2 + \|\mathbf{E}[Sx]\|^2}, \frac{t}{\|S\|_{op}}\right)\right).$$

For the other side, it is analogous.

Proof As remarked in the proof of Lemma 14, combining Theorem 12 from Anari et al. (2021) and Proposition 2.18 from Sambale and Sinulis (2019) yields that our model satisfies Glauber-MLSI(ρ), $\mathfrak{d}^+$-MLSI(2ρ) and Po(2ρ). Thus, we can use the second form of Corollary 1.2 in Sambale and Sinulis (2019):

Let $f(x) = x^\top Sx$. By the proof of Lemma 2.17 in Sambale and Sinulis (2019), we know that we can take $\Gamma = \mathfrak{d}^+$, $g(x) = 4\|Sx\|$, and $b = 8\|S\|_{op}$. It now remains to upper bound $\mathbf{E}[\|SX\|])^2$. To do that, we can again use the Poincaré inequality (Lemma 11).

$$\begin{aligned}
\mathbf{E}[\|SX\|]^2 &\leq \mathbf{E}\left[\|SX\|^2\right]) \\
&= \sum_{i=1}^n \mathbf{E}[(S_i^\top X)^2] \\
&= \sum_{i=1}^n \left(\mathbf{E}[S_i^\top X]^2 + \mathbf{Var}[S_i^\top X]\right) \\
&= \|\mathbf{E}[SX]\|_2^2 + \sum_{i=1}^n \mathbf{Var}[S_i^\top X] \\
&\leq \|\mathbf{E}[SX]\|_2^2 + 2\rho\sum_{i=1}^n \|S_i\|_2^2 \\
&= \|\mathbf{E}[SX]\|_2^2 + 2\rho\|S\|_F^2
\end{aligned}$$

By substituting this upper bound into (18), and notice that $\|S\|_{op} \leq \|S\|_F$ and $\rho \geq 1$ by assumption, we obtain the desired inequality. ■

Having obtained this concentration result, showing that the second derivative is large with high probability boils down to the following two tasks.

(a) First, we need to establish that the second derivative is lower bounded by a degree 2 polynomial. To do this, we need to show that the terms $\operatorname{sech}(J_i X)^2$ are lower bounded by a constant with high probability.

(b) Second, we need to show that the degree 2 polynomial that lower bounds the second derivative is large enough on expectation.

We next show how to establish each of these two properties. We start by addressing (b), namely establishing a lower bound for the expectation of the second moment part of the second derivative, which will prove useful later. The proof involves using the Hubbard-Stratonovich transform to

decompose the Ising model into a mixture of product measures. For each product measure, lower-bounding the second moment is a much simpler problem. However, some product measures in the decomposition will have large external fields, giving weak lower bounds. We use properties of this decomposition to establish that with at least a constant probability, the external field in the decomposition will be bounded.

Lemma 22 *Let $x \sim \mu$ be an Ising model satisfying Assumption 17. Let A be any matrix. Then, we have the following anti-concentration bound:*

$$\mathbf{E}(x^\top A^\top Ax) - \|\mathbf{E}[Ax]\|^2 \geq \frac{1}{2}\left(1 - \tanh^2(4\lambda\sqrt{\rho})\right)\|A\|_F^2 .$$

Proof We use the decomposition described in (14) and (13). In particular, let π be the distribution of vector y and $\mathbf{Pr}_{0,J^*x}(x) \sim \exp(\langle J^*y, x\rangle)$ the corresponding product measure. We can have that

$$\mathbf{E}(x^\top A^\top Ax) - \|\mathbf{E}[Ax]\|^2 = \sum_{i=1}^n \mathbf{Var}(A_i x) \geq \sum_{i=1}^n \mathop{\mathbf{E}}_{y\sim\pi} \mathop{\mathbf{Var}}_{x\sim\nu(y)}(A_i x) = \sum_{i=1}^n \sum_{j=1}^n \mathop{\mathbf{E}}_{y\sim\pi} A_{ij}^2(\operatorname{sech}^2((J^*y)_j))$$

We know that $(J^*y) = (J^*x) + ((J^*)^{1/2}g)$ is a multivariate Gaussian distribution with means J^*x and covariance matrix J^*. So for each (J^*y), we have mean J_i^*x and variance is J_{ii}^*. Therefore, our plan is to prove that J_i^*x and $\mathcal{N}(0, J_{ii}^*)$ are small with high probability.

Consider P to be the Glauber Dynamics of the Ising model. Consider $f(x) = \sum_{j=1}^n J_{ij}^* x_j$. So, because of the distribution, we know that J_i^*x has zero mean. Also, we can calculate that

$$v(f) \leq \frac{1}{n}\sum_{i=1}^n 4(J_{ij}^*)^2 \leq \frac{4\lambda^2}{n}.$$

We know that μ satisfies Glauber-MLSI(ρ). Therefore, by Lemma 16, we can bound the probability that y_j is large:

$$\begin{aligned}&\mathbb{P}((J^*y)_j > a) \leq \mathbb{P}(f(x) > a/2) + \mathbb{P}(\mathcal{N}(0, J_{ii}) > a/2)\\ \leq &\exp(-\frac{a^2}{8\rho\lambda^2}) + \exp(-\frac{a}{8}) \leq 2\exp(-\frac{a^2}{8\rho\lambda^2}).\end{aligned}$$

Therefore, we take $a = 4\sqrt{\rho}\lambda$, we can get that the probability that $(J^*y)_j \leq 4\sqrt{\rho}\lambda$ is at least $1 - 2/e^2 > 1/2$. Therefore, we have

$$\sum_{i=1}^n \mathbf{Var}(A_i x) = \sum_{i=1}^n \sum_{j=1}^n \mathop{\mathbf{E}}_{y\sim\pi} a_{ij}^2(1 - \tanh^2((J^*y)_j)) \geq \frac{1}{2}\left(1 - \tanh^2(4\lambda\sqrt{\rho})\right)\|A\|_F^2 .$$

■

We are now ready to prove our main Lemma about lower-bounding the second derivative of the pseudolikelihood. The proof essentially involves addressing (a), i.e., showing that $\operatorname{sech}(J_i X)^2$ is lower bounded by a constant with high probability. Since we only know a bound on $\|J\|_2$, in general, this quantity could be very small. We use concentration results for second-degree polynomials to relate $\operatorname{sech}(J_i X)$ with $\operatorname{sech}(J_i^* X)$, which does not depend on the matrix direction J. The result is stated for one sample for simplicity, but we'll see how to apply it for multiple samples in the sequel.

Lemma 23 *Suppose $X \in \{-1,1\}^n$ is a sample drawn from an Ising model satisfying Assumption 17. Let $A \in \mathcal{S}_0^n$ be a symmetric matrix with $\sqrt{\|A\|_F^2 + \|\mathbf{E}[Ax]\|^2} \leq M$ for some constant $M > 0$. Let us also denote $J' = J^* + A$. Then, for any J that lies in the line segment connecting J^* and J', we have*

$$\frac{\partial^2 \phi(J)}{\partial A^2} \geq K_2 \cdot (\|A\|_F^2 + \|\mathbf{E}[Ax]\|^2) \cdot \min_{i \in [n]} \operatorname{sech}^2(|J_i^* X^{(k)}| + h_i + K_1 M)$$

with probability at least

$$1 - \exp(-\frac{c}{\rho^2} \cdot \min(t_1^2, t_1)) - \exp(-\frac{c}{\rho^2} \cdot \min(t_2^2, t_2)),$$

where c is an absolute constant as in Lemma 21, and the expression K_1, K_2 are as follows:

$$K_1 = \sqrt{t_1 + 2\rho}$$

$$K_2 = \frac{1}{2}\big(1 - \tanh^2(4\lambda\sqrt{\rho})\big) - t_2$$

Proof We have that

$$\begin{aligned}
\frac{\partial^2 \phi(J)}{\partial A^2} &= \sum_{i=1}^n (A_i X)^2 \operatorname{sech}^2(J_i X) \\
&\geq \sum_{i=1}^n (A_i X)^2 \operatorname{sech}^2(|J_i^* X| + |(J_i - J_i^*)X|) \\
&\geq \sum_{i=1}^n (A_i X)^2 \operatorname{sech}^2(|J_i^* X| + |A_i X|)
\end{aligned}$$

The last inequality follows since $|(J_i - J_i^*)X| \leq |((J')_i - J_i)X| = |A_i X|$, since J lies in the segment connecting J', J^*. Now, let us use Lemma 21 for the quadratic form $x^\top A^\top A x$. When substituting $t \leftarrow t(\|A\|_F^2 + \|\mathbf{E}[AX]\|^2)$, this gives

$$X^\top A^\top A X - \mathbf{E}[X^\top A^\top A X] = \sum_{i=1}^n (A_i X)^2 - \mathbf{E}[X^\top A^\top A X] \leq t(\|A\|_F^2 + \|\mathbf{E}[AX]\|^2). \quad (19)$$

This bound of X means in particular that

$$\sum_{i=1}^n (A_i X)^2 \leq \mathbf{E}[X^\top A^\top A X] + t(\|A\|_F^2 + \|\mathbf{E}[AX]\|^2).$$

By Lemma 21, this happens with probability at least

$$1 - \exp(-c/\rho^2 \min(\frac{t^2(\|A\|_F^2 + \|\mathbf{E}[AX]\|^2)^2}{\|A^\top A\|_F^2 + \|\mathbf{E}[A^\top A X]\|^2}, \frac{t(\|A\|_F^2 + \|\mathbf{E}[AX]\|^2)}{\|A^\top A\|_{op}})).$$

Using the well known properties $\|A^\top A\|_{op} = \|A\|_{op}^2$, $\|\mathbf{E}[A^\top AX]\|^2 = \|A^\top \mathbf{E}[AX]\|^2 \leq \|A\|_{op}\|\mathbf{E}[AX]\|^2$ and $\|A^\top A\|_F \leq \|A\|_F\|A\|_{op}$, we can lower bound the probability by

$$\begin{aligned}
&1-\exp(-c/\rho^2\min(\frac{t^2(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)^2}{\|A^\top A\|_F^2+\|\mathbf{E}[A^\top AX]\|^2},\frac{t(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)}{\|A^\top A\|_{op}}))\\
\geq&1-\exp(-c/\rho^2\min(\frac{t^2(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)^2}{\|A\|_{op}^2(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)},\frac{t(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)}{\|A\|_{op}^2}))\\
=&1-\exp(-c/\rho^2\frac{\|A\|_F^2+\|\mathbf{E}[AX]\|^2}{\|A\|_{op}^2}\min(t,t^2)).
\end{aligned}$$

Now, to finally deal with upper bound of $\mathbf{E}[X^\top A^\top AX]$, we use lemma 15 (since μ satisfy Glauber-MLSI(ρ), hence also Po(2ρ)) to decude that, if (19) holds, then.

$$\begin{aligned}
&\sum_{i=1}^n (A_iX)^2 \leq \mathbf{E}[X^\top A^\top AX]+t(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)\\
\leq&(1+t)(\|\mathbf{E}[AX]\|^2)+(t+2\rho)\|A\|_F^2 \leq (t+2\rho)(\|\mathbf{E}[AX]\|^2+\|A\|_F^2) \leq (t+2\rho)M^2.
\end{aligned}$$

We know that (19) holds with probability at least $1-\exp(-\frac{c(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)}{\rho^2\|A\|_{op}^2}\cdot\min(t^2,t))$, thus the same probability that the above chain of probability holds. Denote $K_1=\sqrt{t+2\rho}$. It follows that with the same probability we have $|A_iX|\leq K_1M$ for all i. Thus, we have that

$$\begin{aligned}
\frac{\partial^2\phi(J)}{\partial A^2} &\geq \sum_{i=1}^n (A_iX)^2\operatorname{sech}^2(|J_i^*X|+|A_iX|)\\
&\geq \sum_{i=1}^n (A_iX)^2\operatorname{sech}^2(|J_i^*X|+K_1M)\\
&\geq \min_{i\in[n]}\operatorname{sech}^2(|J_i^*X|+K_1M)\sum_{i=1}^n (A_iX)^2
\end{aligned}$$

We now need to lower bound $\mathbf{E}[X^\top A^\top AX]$ in order to get a high probability lower bound for $\sum_{i=1}^n (A_iX)^2$. Using Lemma 21 we have

$$X^\top A^\top AX \geq \mathbf{E}[X^\top A^\top AX]-t(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)$$

holds for at least $1-\exp(-\frac{c(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)}{\rho^2\|A\|_{op}^2}\cdot\min(t^2,t))$ probability. Thus, by Lemma 22, if the above inequality holds, we can deduce that

$$\begin{aligned}
X^\top A^\top AX &\geq \mathbf{E}[X^\top A^\top AX]-t(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)\\
&\geq (\frac{1}{2}(1-\tanh^2(4\lambda\sqrt{\rho}))-t)(\|A\|_F^2+\|\mathbf{E}[AX]\|^2)
\end{aligned}$$

Denote $K_2=\frac{1}{2}\big(1-\tanh^2\big(4\lambda\sqrt{\rho}\big)\big)-t_2$. Using the union bounds, and noticing that $\|A\|_F > \|A\|_{op}$, we can finish the proof. ■

We now use Lemma 23 to lower-bound the second derivative when we have multiple independent samples. The main issue is that we need to choose the parameters t_1, t_2 to ensure K_2 is bounded away from 0, while at the same time obtaining a probability of failure that decays exponentially with the number of samples l. As we see, the particular block structure of the matrix in that case is crucial for the argument to go through.

Theorem 24 *Suppose $X^{(1)}, \dots, X^{(l)} \in \{-1,1\}^n$ are independent samples from an Ising model satisfying Assumption 17. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with $\sqrt{\|A\|_F^2 + \|\mathbf{E}[Ax]\|^2} \le M \le 1$ for some constant $0 < M < 1$. Let us also denote $J_1 = J^* + A$. Then, for any J that lies in the line segment connecting J^* and J_1, we have*

$$\frac{\partial^2 \phi(J^{(l)})}{\partial (A^{(l)})^2} \ge r(\lambda,\rho)\cdot l \cdot (\|A\|_F^2 + \|\mathbf{E}[Ax]\|^2) \cdot \min_{k\in[l], i\in[n]} \operatorname{sech}^2(|J_i^* X_i^{(k)}| + K(\lambda,\rho)M + h_i)$$

with probability at least $1 - \exp(-C(\lambda,\rho)\cdot l)$, where $r(\lambda,\rho), C(\lambda,\rho)$ are constants that depend only on λ, ρ. More specifically,

$$r(\lambda,\rho) = C(\lambda,\rho)^2/4; \quad C(\lambda,\rho) = \frac{c}{48\rho^2}\cdot \operatorname{sech}(4\lambda\sqrt{\rho})^4; \quad K(\lambda,\rho) = \frac{10\rho^{5/4}\sqrt{\lambda}}{c}$$

where c is the constant defined in Lemma 21.

Proof Using the block structure of the matrix, we can write

$$\frac{\partial^2 \phi(J^{(l)})}{\partial (A^{(l)})^2} = \sum_{k=1}^{l} \underbrace{\sum_{i=1}^{n} (A_i X^{(k)})^2 \operatorname{sech}^2(J_i X^{(k)} + h_i)}_{G_k}$$

The random variables $G_1, \dots, G_k$ are independent and identically distributed. We use Lemma 23 with the substitution

$$t_2 = \frac{\operatorname{sech}(4\lambda\sqrt{\rho})^2}{4}, \quad t_1 = \frac{\log(\frac{2\rho^2}{ct_2^2})\rho^2}{c}.$$

We can clearly see that $t_2 < 1/4$ and $t_1 > 1$ (we, without loss of generality, take the absolute constant c in Lemma 21 smaller than 1). Therefore, we have

$$\begin{aligned}
&1 - \exp(-c/\rho^2 \cdot \min(t_1^2, t_1)) - \exp(-c/\rho^2 \min(t_2^2, t_2)) \\
=&1 - \exp(-c/\rho^2 \cdot t_1) - \exp(-c/\rho^2 \cdot t_2^2) \\
=&1 - \exp(-c/\rho^2 \cdot \frac{\log(\frac{2\rho^2}{ct_2^2})\rho^2}{c}) - \exp(-c/\rho^2 \cdot t_2^2) \\
=&1 - c/\rho^2 \cdot t_2^2/2 - \exp(-c/\rho^2 \cdot t_2^2)
\end{aligned}$$

We know that for small $x < 1/4$, $1 - e^{-x} - x/2 \ge x/3$. So, for all k, with probability at least

$$p := c/\rho^2 \cdot t_2^2/3 = c/\rho^2 \cdot \frac{\operatorname{sech}(4\lambda\sqrt{\rho})^4}{16} \cdot \frac{1}{3} \geq c/\rho^2 \cdot \operatorname{sech}(4\lambda\sqrt{\rho})^4/48,$$

the following holds

$$G_k \geq K_2 \cdot (\|A\|_F^2 + \|\mathbf{E}[Ax]\|^2) \cdot \min_{i\in[n]} \operatorname{sech}^2\left(|J_i^* X^{(k)}| + K_1 M + h_i\right),$$

where K_1, K_2 are defined in Lemma 23. We have

$$K_2 = \frac{1}{2}\operatorname{sech}^2(4\lambda\sqrt{\rho}) - t_2 \geq \operatorname{sech}^2(4\lambda\sqrt{\rho})/4,$$

and

$$K_1 = \sqrt{t_1 + 2\rho} \leq 2\sqrt{t_1} = \frac{2\rho}{\sqrt{c}} \cdot \sqrt{\log(\frac{2\rho^2}{ct_2^2})} = \frac{2\rho}{\sqrt{c}} \cdot \sqrt{\log(\frac{32\rho^2}{c}) + 4\log(\cosh(4\lambda\sqrt{\rho}))} \leq \frac{10\rho^{5/4}\sqrt{\lambda}}{c}.$$

We will now use the independence of the different samples. Let's call the above event E_k, then $\mathbf{1}\{E_k\}$ is a Bernoulli random variable with probability at least p. By the standard Chernoff bound for Bernoulli random variables, we have that for all $\delta \in (0,1)$

$$\mathbf{Pr}\left[\sum_{k=1}^{l} \mathbf{1}\{E_k\} \leq (1-\delta)lp\right] \leq \exp(-\delta^2 lp/2),$$

Choosing $\delta = 1/2$ yields that with probability at least $1 - \exp(-lp/8)$ at least an $lp/4$ fraction of the events E_k will be satisfied. We conclude that with the same probability

$$\begin{aligned}\frac{\partial^2\phi(J^{(l)})}{\partial(A^{(l)})^2} \geq& \sum_{k:\mathbf{1}\{E_k\}=1} G_k \\ \geq& p^2/4 \cdot l \cdot (\|A\|_F^2 + \|\mathbf{E}[Ax]\|^2) \cdot \min_{k\in[l],i\in[n]} \operatorname{sech}^2\left(|J_i^* X_i^{(k)}| + K_1 M + h_i\right)\end{aligned}$$

Plugging in p, K_1, M yields the result. ■

B.3. Learning in Frobenius norm

After establishing the concentration of the first derivative and anti-concentration of the second derivative, we now combine them to show that we can learn in the Frobenius norm.

The precise statement is given in the following lemma. Its purpose is to show that if for some matrix J_1 the norm $\|J^* - J_1\|_F$ is large, then with high probability J_1 will have a lower pseudo-likelihood value than J^*, which means that it will not be selected as the maximizer of the pseudolikelihood. The proof essentially combines the concentration and anticoncentration properties of the two derivatives for a single matrix J_1.

Lemma 25 *Let X be a sample drawn from an Ising model with interaction matrix J^* satisfying Assumption 17 and let $J_1 \in \mathcal{R}$ be a different matrix, with $A = J_1 - J^*$. Assume $\sqrt{\|A\|_F^2 + \|\mathbf{E}[Ax]\|^2} \leq M < 1$, for some $M > 0$. Then, there exist constants C, C', r, K that depend only on ρ, λ, such that for any $t > 0$, with probability at least*

$$1 - \exp(-C \cdot l) - 8/3 \exp\left(-C' \cdot l \cdot \min(t^2, t) \cdot M^2\right),$$

holds

$$\phi(J_1) \geq \phi(J^*) + r \cdot l \cdot M^2 \cdot \min_{k \in [l], i \in [n]} \operatorname{sech}^2(|J_i^* X_i^{(k)}| + KM + h_i) - t \cdot l \cdot M^2.$$

Here, C, r, K are the constants $C(\rho, \lambda), r(\rho, \lambda), K(\rho, \lambda)$ defined in Theorem 24, and $C' = \frac{1}{2^{21}\rho^2\lambda^4}$ is the constant defined in Theorem 18.

Proof [Proof of Lemma 25] We begin as in Lemma 1 in Dagan et al. (2021) by defining the function $J : [0, 1] \mapsto \mathbb{R}^{n \times n}$ as $J(t) = (1 - t)J^* + tJ_1$. Let $A = J_1 - J^*$. By definition, then we have

$$\frac{d\phi(J(t))}{dt} = \left.\frac{\partial \phi(J)}{\partial A}\right|_{J=J(t)} \quad , \quad \frac{d^2\phi(J(t))}{dt^2} = \left.\frac{\partial^2 \phi(J)}{\partial A^2}\right|_{J=J(t)} .$$

Now define

$$\alpha = \min_{t \in [0,1]} \frac{d^2\phi(J(t))}{dt^2} \quad , \quad \gamma = \left.\frac{d\phi(J(t))}{dt}\right|_{t=0}$$

By Taylor's theorem, we have that

$$\phi(J(1)) \geq \phi(J(0)) + \gamma + \frac{\alpha}{2}$$

By the definition of $J(t)$, this is equivalent to

$$\phi(J_1) \geq \phi(J^*) + \gamma + \frac{\alpha}{2} .$$

Now, using Theorem 24 immediately gives

$$\alpha \geq r \cdot l \cdot \|A\|_F^2 \cdot \min_{k \in [l], i \in [n]} \operatorname{sech}^2(|J_i^* X_i^{(k)}| + KM)$$

with probability at least $1 - \exp(-C \cdot l)$, where r, C, K are constants that depend only on ρ, λ, as in Theorem 24. At the same time, we can apply Lemma 18, where A is now $A^{(l)}$ and so instead of $t \leftarrow t(\|A\|_F^2 + \|\mathbf{E}[Ax]\|^2)$ we have $t \leftarrow tlM^2$. After replacing t with $t\|A\|_F$ and notice that $\|A\|_{op} \leq \|A\|_F \leq 1$ we obtain

$$\mathbf{Pr}[|\gamma| > tlM^2] \leq \frac{8}{3} \exp\left(-C' \cdot l \cdot \min(t, t^2) \cdot M^2\right)$$

Where C' is the constant in Lemma 19. Thus, with probability at least

$$1 - \exp(-C \cdot l) - 8/3 \exp\left(-C' \cdot l \cdot \min(t, t^2) \cdot M^2\right)$$

we have that

$$\phi(J_1) \geq \phi(J^*) + r \cdot l \cdot M^2 \cdot \min_{k\in[l], i\in[n]} \operatorname{sech}^2(|J_i^* X_i^{(k)}| + KM) - t \cdot l \cdot M^2.$$

And the proof is complete. ■

We now have to establish a similar property as in Lemma 25 but across all directions J_1 with high probability. If we had this property, we could conclude that whatever matrix is returned from maximizing the pseudolikelihood function, it needs to be close to the true matrix J^* in Frobenius norm. To do this, we utilize the fact that ϕ is Lipschitz with respect to the spectral norm.

Lemma 26 (Lemma 10, Dagan et al. (2021)) *For any symmetric matrices $J_1, J_2 \in \mathbb{R}^{n\times n}$, and for l samples,*

$$|\phi(J_1) - \phi(J_2)| \leq n \cdot l \cdot \|J_1 - J_2\|_{op}$$

We next define the important notion of an ε-net.

Definition 27 *Given a metric space $(\mathcal{U}, d)$ and $\varepsilon > 0$, we say that a subset $\mathcal{N} \subseteq \mathcal{U}$ is an ε-*net *for $\mathcal{U}$ if for every $u \in \mathcal{U}$ there exists a $v \in \mathcal{N}$ such that $d(u, v) \leq \varepsilon$. The cardinality of the smallest possible ε-net is denoted by $\mathcal{N}(\mathcal{U}, d, \varepsilon)$. We also refer to $\mathcal{N}(\mathcal{U}, d, \varepsilon)$ as the ε-covering number of the set $\mathcal{J}$.*

The strategy now will be to show that all matrices that are far from J^* in Frobenius norm will have a higher pseudolikelihood value with high probability. Arguing simultaneously over all such matrices is a daunting task, since this is an infinite set. The strategy that was employed in Dagan et al. (2021) was the following: first, we construct an ε-net to cover the entire space of matrices. By choosing ε sufficiently small and using the Lipschitzness property, we can show that any point in the space has a pseudolikelihood value close to some point in the net. Consequently, if we can guarantee that with high probability all points in the net have pseudolikelihood value smaller than that of J^*, then the same should be true for all points in the set as well.

However, this approach cannot work in our case, because by Lemma 25 we can only argue about the value of ϕ for points that are close to J^* in Frobenius norm. Thus, our goal will be to show that there is a *shell* of matrices of the form $\{J : \varepsilon \leq \|J^* - J\|_F \leq M\}$, such that all points in this shell have a higher pseudolikelihood value than J^*. It will then follow from the convexity of ϕ that the same is true for points outside of this shell as well. It would then follow that the true minimizer $\hat{J}$ should satisfy $\|\hat{J} - J^*\|_F \leq \varepsilon$, which is our final estimation bound. The challenge is to show that for ε taking a relatively small value, this property will hold, as this will result in a small estimation error. To argue about that, we are aided by the fact that we can choose l large enough to make ε smaller than M, so this shell is not empty. The details are given below.

Theorem 28 *Let $X_1, \ldots, X_l$ be independent samples drawn from an Ising model with interaction matrix J^* satisfying Assumption 17. Let $\hat{J} \in \mathbb{R}^{n\times n}$ be the estimate of J^* that is obtained by maximizing the pseudo-likelihood function (6), i.e.*

$$\hat{J}^{(l)} := \operatorname*{argmax}_{J^{(l)}\in\mathcal{R}^{(l)}} \phi(J; X) \ .$$

Then, for any $\varepsilon, \delta \in (0,1)$*, if*

$$l \geq \exp\left(C_0\left(\lambda^2\rho + \lambda\sqrt{\rho}\sqrt{\log(n/(\delta\varepsilon))} + \rho^{5/4}\sqrt{\lambda}\varepsilon + h_{\max}\right)\right) \cdot \frac{n^2}{\varepsilon^2},$$

with probability at least $1-\delta$*,*

$$\sqrt{\|J - J^*\|_F^2 + \|\mathbf{E}[(J-J^*)X]\|^2} \leq \varepsilon$$

holds. Here, C_0 *is a universal constant.*

Before we go to the proof, we should clarify that $\hat{J}$ may not satisfy MLSI(ρ). We only need the fact that at the ground truth J^*, the MLSI satisfies. If $\hat{J}$ satisfies MLSI(ρ) (and moreover for all $t \in [0,1]$, $tJ^* + (1-t)\hat{J}$, if it satisfies MLSI(ρ)), the bound of TV distance from KL divergence may be better (see Lemmas 39 and 40.)

Proof Let $\varepsilon \in (0, 1/2)$ and set $M = 2\varepsilon$. Let us define the shell

$$\mathcal{R}_\varepsilon := \{J \in \mathcal{R} : \varepsilon \leq \sqrt{\|J - J^*\|_F^2 + \|\mathbf{E}[(J-J^*)X]\|^2} \leq 2\varepsilon\}$$

Our goal will be to choose l such that with high probability $\hat{J} \notin \mathcal{R}_\varepsilon$. First of all, since $\mathcal{R}_\varepsilon \subseteq \mathcal{R}$, we have that for any $\theta > 0$

$$\mathcal{N}(\mathcal{R}_\varepsilon, \|\cdot\|_{op}, \theta) \leq \mathcal{N}(\mathcal{R}'_\varepsilon, \|\cdot\|_{op}, \theta)$$

Here, we have defined the set $\mathcal{R}'_\varepsilon := \{J \in \mathcal{R} : \|J - J^*\|_F \leq 2\varepsilon\}$.

Thus, it suffices to bound $\mathcal{N}(\mathcal{R}, \|\cdot\|_{op}, \theta)$. To that end, we can view $\mathcal{R}$ as a subset of n^2-dimensional Euclidean space, where the basis vectors are matrices $\{E_{ij}\}_{i,j=1}^n$, where E_{ij} has the i,j entry 1 and the rest 0. Thus, we seek to cover the ball of matrices with spectral radius at most 1 with balls of radius θ. Notice that the ball with spectral radius θ is contained inside the ball with Frobenius radius θ. Since the ball of spectral radius 1 is a centrally symmetric convex body, we can apply Corollary 4.1.15 from Artstein-Avidan et al. (2021), which is based on a standard volume argument, to obtain

$$\mathcal{N}(\mathcal{R}_\varepsilon, \|\cdot\|_{op}, \theta) \leq \mathcal{N}(\mathcal{R}'_\varepsilon, \|\cdot\|_{op}, \theta) \leq \mathcal{N}(\mathcal{R}'_\varepsilon, \|\cdot\|_F, \theta) \leq \left(1 + \frac{2\varepsilon}{\theta}\right)^{n^2} \tag{20}$$

Let $\mathcal{U} \subseteq \mathcal{R}$ be an θ/n-net of $\mathcal{R}_\varepsilon$ of cardinality $\mathcal{N}(\mathcal{R}_\varepsilon, \|\cdot\|_{op}, \theta/n)$, where θ will be chosen in the sequel. Then, applying Lemma 25 and a union bound gives that for all $t > 0$, with probability at least

$$1 - \mathcal{N}(\mathcal{R}'_\varepsilon, \|\cdot\|_{op}, \theta/n)\left(\exp(-C \cdot l) + 8/3\exp\left(-C' \cdot l \cdot \min(t^2, t) \cdot \varepsilon^2\right)\right).$$

We have that the following event occurs.

$$\mathcal{E} := \{\phi(J) \geq \phi(J^*) + l \cdot (r \cdot \min_{k\in[l], i\in[n]} \operatorname{sech}^2(|J_i^* X_i^{(k)}| + 2K\varepsilon + h_i) - 4t) \cdot M^2 \quad , \forall J \in \mathcal{U}\} \ .$$

Here, C, r, K are defined in Theorem 24, and C' is defined in Theorem 18. Now let us assume $\mathcal{E}$ happens. We now upper bound $|J_i^* X^{(k)}|$. Indeed, using Lemma 16 we have that for each k

$$\mathbf{Pr}[|J_i^* X^{(k)}| \geq a] \leq \exp\left(-\frac{1}{8\rho\lambda^2}a^2\right)$$

By union bound over all $i \in [n], k \in [l]$, we have

$$\mathbf{Pr}[\max_{i\in[n],k\in[l]} |J_i^* X^{(k)}| \geq a] \leq l \cdot n \cdot \exp\left(-\frac{1}{8\rho\lambda^2}a^2\right)$$

Thus, by choosing $a = C''\lambda\sqrt{\rho\log((nl)/\delta)}$ for some suitable absolute constant C'', we get that with probability at least $1-\delta/2$ (we remind the readers that $h_{\max}$ is the maximum value of $|h_i|$)

$$\min_{k\in[l],i\in[n]} \operatorname{sech}^2(|J_i^* X_i^{(k)}| + 2K\varepsilon) \geq \operatorname{sech}^2\left(C''\lambda\sqrt{\rho\log(nl/\delta)} + 2K\varepsilon + h_{\max}\right) := \xi(n)^{-1} \ .$$

We notice that $\xi(n)$ grows sub-polynomially, i.e. $\xi(n) = o(n^r)$ for any $r > 0$. Thus, we conclude that with probability at least

$$1 - \mathcal{N}(\mathcal{R}'_\varepsilon, \|\cdot\|_{op}, \theta/n)\left(\exp(-C\cdot l) + 8/3\exp\left(-C'\cdot l\cdot\min(t^2,t)\cdot\varepsilon^2\right)\right) - \frac{\delta}{2},$$

we have that

$$\phi(J) \geq \phi(J^*) + r\cdot l\cdot\varepsilon^2\cdot\xi(n)^{-1} - 4t\cdot l\cdot\varepsilon^2 \quad ,\forall J\in\mathcal{U} \ .$$

Now, choosing $t \leq r\xi(n)^{-1}/8$. Notice that r is really small compare to C and C', we have $\exp(-C\cdot l) \leq \exp\left(-C'\cdot l\cdot\min(t^2,t)\cdot\varepsilon^2\right) = \exp\left(-C'\cdot l\cdot t^2\cdot\varepsilon^2\right)$ and we have that with probability at least

$$1 - \frac{11}{3}\mathcal{N}(\mathcal{R}'_\varepsilon, \|\cdot\|_{op}, \theta/n)\exp\left(-C'\cdot r^2\cdot\xi(n)^{-2}\cdot l\cdot\varepsilon^2/64\right) - \frac{\delta}{2},$$

the following event holds

$$\mathcal{E}' := \left\{\phi(J) \geq \phi(J^*) + \frac{1}{2}r\cdot l\cdot\varepsilon^2\cdot\xi(n)^{-1} \quad ,\forall J\in\mathcal{U}\right\} \ .$$

Let us now see how we should choose l so that

$$\mathcal{N}(\mathcal{R}'_\varepsilon, \|\cdot\|_{op}, \theta/n)\exp\left(-C'\cdot r^2\cdot\xi(n)^{-2}\cdot l\cdot\varepsilon^2/64\right) < \frac{\delta}{2} \ .$$

Using the covering number bound (20), it suffices to pick

$$l \geq \frac{64\xi(n)\left(n^2\log(1+2\varepsilon n/\theta) + \log(10/\delta)\right)}{C'r^2\varepsilon^2} \tag{21}$$

Finally, let us see how to choose θ. We would like to show that, if event $\mathcal{E}'$ holds, then for an arbitrary element $J \in \mathcal{R}_\varepsilon$ we have $\phi(J) > \phi(J^*)$. By definition, there exists $\overline{J} \in \mathcal{U}$ with $\|J - \overline{J}\|_{op} \leq \theta/(2n)$. This, combinined with Lemma 26, implies that

$$\phi(J) \geq \phi(\overline{J}) - \frac{\theta}{2} \geq \phi(J^*) + \frac{1}{2}r\cdot l\cdot\varepsilon^2\cdot\xi(n)^{-1} - \frac{\theta}{2}\cdot l$$

The last quantity is $> \phi(J^*)$ if we pick $\theta = \frac{1}{2}r\cdot\varepsilon^2\cdot\xi(n)^{-1}$. Thus, (21) becomes

$$l \geq \frac{\xi(n)\left(n^2\log(1+2n\cdot\xi(n)/(r\varepsilon)) + \log(10/\delta)\right)}{C'r^2\varepsilon^2} \tag{22}$$

Thus, with this choice of l, we know that with probability at least $1-\delta/2$

$$\phi(J) > \phi(J^*) \quad ,\forall J \in \mathcal{R}_\varepsilon \ .$$

We now prove that if we select

$$l \geq \exp\left(C_0\left(\lambda^2\rho + \lambda\sqrt{\rho}\sqrt{\log(n/(\delta\varepsilon))} + \rho^{5/4}\sqrt{\lambda}\varepsilon + h_{\max}\right)\right)\cdot\frac{n^2}{\varepsilon^2},$$

for some absolute constant $C_0 > 1$, the inequality (22) holds. First, we know that, for any positive number A, B, D, if $X \geq 64(A^2 + A\sqrt{B} + D)$, we have $X/2 \geq 4(A\sqrt{X+B} + D)$. This is because

$$X/2 \geq X/4 + 16(A^2 + A\sqrt{B} + D) \geq 4A\sqrt{X} + 16A\sqrt{B} + 16D \geq 4(A\sqrt{X+B} + D).$$

Consider

$$A = C''\lambda\sqrt{\rho},\ B = \log(n^3/(\delta\varepsilon^2)),\ D = 2K\varepsilon + h_{\max} + \log(\frac{\log(1+2n/(r\varepsilon)) + \log(10/\delta)}{C'r^2}).$$

Let $X = 64(A^2 + A\sqrt{B} + D)$. So, we can deduce that, if $l \geq \exp(64(A^2 + A\sqrt{B} + D))\cdot\frac{n^2}{\varepsilon^2}$, we have

$$\begin{aligned}
\xi(n) &= \operatorname{sech}^2(C''\lambda\sqrt{\rho\log(nl/\delta)} + 2K\varepsilon + h_{\max}) \\
&\leq \exp(2C''\lambda\sqrt{\rho}\sqrt{\log(l\varepsilon^2/n^2) + \log(n^3/(\delta\varepsilon^2))} + 2(2K\varepsilon + h_{\max})) \\
&\leq \exp(2A\sqrt{\log(l\varepsilon^2/n^2) + B} + 2D) \\
&\leq \exp(2A\sqrt{X+B} + 2D) \leq \exp(X/4).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
l &\geq \frac{n^2}{\varepsilon^2}\cdot\exp(64(A^2 + A\sqrt{B} + D)) \geq \frac{n^2}{\varepsilon^2}\cdot\exp(X) = \frac{n^2}{\varepsilon^2}\cdot\exp(X/4)^2\cdot\exp(X/2) \\
&\geq \frac{n^2}{\varepsilon^2}\cdot\xi(n)^2\cdot\exp(D) \geq \frac{n^2}{\varepsilon^2}\cdot\xi(n)^2\cdot(\frac{\log(1+2n/(r\varepsilon)) + \log(10/\delta)}{C'r^2}) \\
&\geq \frac{n^2}{\varepsilon^2}\cdot\xi(n)\cdot(\frac{\log(1+2n\xi(n)/(r\varepsilon)) + \log(10/\delta)}{C'r^2}) \\
&\geq \xi(n)\cdot(\frac{n^2\log(1+2n\xi(n)/(r\varepsilon)) + \log(10/\delta)}{C'r^2\varepsilon^2})
\end{aligned}$$

A, B can be seen directly in the bound of l, now we estimate the D. We know that D contains $2K\varepsilon + h_{\max}$, which correspond to $\rho^{5/4}\sqrt{\lambda}\varepsilon + h_{\max}$ in l ($K = O(\rho^{5/4}\sqrt{\lambda})$, according to Theorem 24). The rest of the terms for $\log(\frac{\log(1+2n/(r\varepsilon))+\log(10/\delta)}{C'r^2})$ is upper bounded by a linear combination of $\log\log(n)$, $\log\log(1/\delta)$, $\log\log(1/\varepsilon)$ and $\log(1/C')$, $\log(1/r)$, where $\log(1/C') = O(\log(\rho) + \log(\lambda))$ is $\log(1/r) = O(\lambda\sqrt{\rho}\cdot\log\rho)$ according to Theorem 24. The previous three terms are absorbed by $\lambda\sqrt{\rho}\sqrt{\log(n/(\delta\varepsilon))}$, and the last two is absorbed by $\lambda^2\rho$. Hence, we can prove that this l satisfies the condition of the inequality (22).

We now go back to the main part. We call the event $\phi(J) > \phi(J^*), \forall J \in \mathcal{R}_\varepsilon$ as $\mathcal{E}''$. We argue that if $\mathcal{E}''$ holds, then for all J with $\|J - J^*\|_F > 2\varepsilon$ we have $\phi(J) > \phi(J^*)$. Indeed, for any

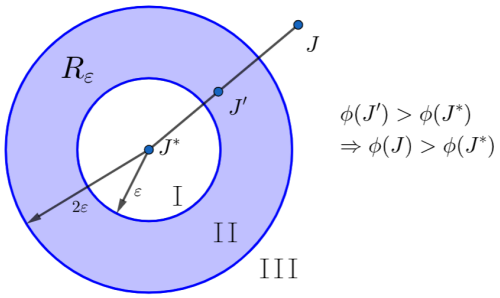

Figure 1: The set $\mathcal{R}_\epsilon$ naturally partitions the whole space into three regions, I, II, and III. We know that all J in region II satisfy $\phi(J) > \phi(J^*)$. If we consider an arbitrary J in region III, then by appropriate scaling, we can find J' in region II that lies on the line connecting J and J^*. The assumption about J^* combined with the convexity of the pseudolikelihood ϕ allows us to conclude that $\phi(J) > \phi(J^*)$. Thus, $\hat{J}$ lies in region I.

such J, the line segment connecting J to J^* intersects $\mathcal{R}_\varepsilon$ in at least one point, call it J'. This is because the inequalities defining $\mathcal{R}_\varepsilon$ scale by a constant as we move from J to J^* (see Figure 1 for an illustration).

Thus, for some $t \in (0,1)$ we can write $J' = (1-t)J^* + tJ$. Now, since ϕ is a convex function, it holds

$$\phi(J') \leq (1-t)\phi(J^*) + t\phi(J) < (1-t)\phi(J') + t\phi(J)$$

The last inequality holds by definition of event $\mathcal{E}''$. By rearranging we get $\phi(J) > \phi(J^*)$. Thus, $\hat{J}$ can only lie inside the Frobenius norm sphere of radius ε around J^*, which concludes the proof. ■

Appendix C. Learning Ising Models with Bounded Width

In this section, our goal will be to establish sample complexity guarantees for learning Ising Models of bounded width in TV distance. We say an Ising Model has width bounded by $M > 0$, if and only if $\|J^*\|_\infty \leq M$. This enables us to handle the second derivative more easily since the sech terms are always lower bounded by a constant that depends on M.

Assumption 29 *In this section, we assume $h = 0$ and $\mathcal{R} = \{J^* \in \mathcal{S}_0^n : \|J^*\|_\infty \leq M\}$ for some $M > 0$.*

For this Section, $\mathcal{R} \subseteq \mathcal{S}_0^n$ will denote the set of matrices with $\|A\|_\infty \leq M$. In Dagan et al. (2021), it was established that we can learn the interaction matrix in the Frobenius norm. Here, we will need a more refined analysis, which will result in stronger guarantees that will enable us to bound the total variation distance between the estimated and the true model.

For the reader's convenience, we remind some important notation that will be used in this section. For a symmetric matrix $J \in \mathbb{R}^{n\times n}$ and a subset $I \subseteq [n]$, we denote by $J_I \in \mathbb{R}^{|I|\times n}$ the matrix consisting only of the rows of J that are indexed by elements in I. We also denote $J_{II} \in \mathbb{R}^{|I|\times|I|}$

the submatrix with rows and columns indexed by I. For non-square matrices, the Frobenius norm extends in the usual fashion

$$\|J_I\|_F^2 = \sum_{i \in I} \sum_{j=1}^{n} J_{ij}^2.$$

We start by briefly highlighting some of the technical tools used in Dagan et al. (2021), as they will prove useful in our case as well. An important observation is that we can select $O(\log n)$ subsets of nodes, such that each submatrix of J^* satisfies Dobrushin's condition. Furthermore, we require that each node belongs to a constant fraction of these subsets. Formally, the following result was proven in Dagan et al. (2021).

Lemma 30 (Lemma 2 from Dagan et al. (2021)) *Let $J^* \in \mathbb{R}^{n \times n}$ be a symmetric matrix with $\|J^*\|_\infty \leq M$ and let $\eta \in (0, M)$. Then, there exist subsets $I_1, \ldots, I_r$ with $r \leq CM^2 \log n / \eta^2$, such that the following properties hold.*

1. *For all $i \in [n]$*
$$|j \in [r] : i \in I_j| = \left\lceil \frac{\eta r}{8M} \right\rceil \quad .$$

2. *For all $j \in [r]$, $\|J^*_{I_j I_j}\|_\infty \leq \eta$.*

This Lemma will allow us to split the first derivative sum into terms, where each term is a sum over the nodes of each subset. Then, for each subset I_j, by property 2, conditioned on the values X_{-I_j}, the model is in high temperature, so we can apply concentration bounds that are valid in that case. For simplicity, for $j \in [r]$ we use the notations

$$\phi_j(J) = \sum_{k=1}^{l} \sum_{i \in I_j} \left(\log \cosh(J_i X^{(k)}) X_i^{(k)} J_i X^{(k)} + \log 2 \right)$$

$$\frac{\partial \phi_j(J^*)}{\partial A} := \sum_{k=1}^{l} \sum_{i \in I_j} (A_i X^{(k)})(\tanh(J_i^* X^{(k)}) - X_i^{(k)})$$

$$\frac{\partial^2 \phi_j(J)}{\partial A^2} := \sum_{k=1}^{l} \sum_{i \in I_j} \operatorname{sech}^2(J_i X^{(k)})(A_i X^{(k)})^2$$

The above are random variables, but we omit the dependence on X for simplicity. We will argue about each component ϕ_j separately, conditioned on the variables X_{-I_j}. By property 1, we have that

$$\phi(J) = \frac{8M}{\eta r} \sum_{j=1}^{r} \phi_j(J) \tag{23}$$

Our goal will be to show that this bound of the first derivative is of the same order as $l \cdot \mathbf{E}[\|AX\|^2]$ with high probability, which is a quantity independent of the conditioning. To do that, we will bound the deviation between the empirical and true mean of the variable $\| \mathbf{E}[AX^{(k)} | X_{-I}^{(k)}]\|^2$, uniformly over all matrices A with small Frobenius norm. The challenge here is that X_{-I_j} comes from a model at low temperature, so we do not have information about its concentration properties. If

we naively use the Chernoff bound, then the large magnitude of $\|\mathbf{E}[AX^{(k)}|X^{(k)}_{-I}]\|^2$ will incur a large concentration radius, which, combined with a union bound over a high-dimensional subset of matrices, will result in high sample complexity. Instead, we notice that because of the special structure of the random variable, it is enough to upper bound the deviation of the random matrix $\mathbf{E}[X|X^{(k)}_{-I_j}]\,\mathbf{E}[X|X^{(k)}_{-I_j}]^\top$ from its mean, which can be done using matrix concentration results. This avoids the costly union bound over the set of matrices and results in a polynomial reduction in the number of samples required. Let us introduce the set of matrices of small Frobenius norm

$$\mathcal{A}_\epsilon := \{J \in \mathbb{R}^{n\times n} : \|A\|_F \le \epsilon\},$$

The details are given in the following Lemma.

Lemma 31 *Suppose $X^{(1)},\dots,X^{(l)} \in \{-1,1\}^n$ are independent samples from an Ising model with interaction matrix J^* satisfying $\|J^*\|_\infty \le M$ and zero external field. Then, for any $t>0$ we have that with probability at least*

$$1 - 2n\exp\left(-\frac{lt^2/2}{4n^2+2nt/3}\right),$$

the following holds

$$\left|\frac{1}{l}\sum_{k=1}^{l}\|\mathbf{E}[AX^{(k)}|X^{(k)}_{-I}]\|^2 - \mathbf{E}\left[\|\mathbf{E}[AX|X_{-I}]\|^2\right]\right| \le t\|A\|_F^2 \ \ ,\forall A$$

Proof We first notice that we can write

$$\begin{aligned}\|\mathbf{E}[AX|X_{-I}]\|^2 &= \mathbf{E}[X|X_{-I}]^\top A^\top A\,\mathbf{E}[JX|X_{-I}] = \mathrm{Tr}\left(\mathbf{E}[X|X_{-I}]^\top A^\top A\,\mathbf{E}[JX|X_{-I}]\right)\\ &= \mathrm{Tr}(A\,\mathbf{E}[X|X_{-I}]\,\mathbf{E}[X|X_{-I}]^\top A^\top)\ .\end{aligned}$$

Let $S = \mathbf{E}[X|X_{-I}]\,\mathbf{E}[X|X_{-I}]^\top$ be this random matrix and denote by $S^{(k)} := \mathbf{E}[X|X^{(k)}_{-I}]\,\mathbf{E}[X|X^{(k)}_{-I}]^\top$ and $S = \mathbf{E}\left[\mathbf{E}[X|X_{-I}]\,\mathbf{E}[X|X_{-I}]^\top\right]$ the l independent samples from the distribution of S. Then, we can write the difference between the empirical and the true mean as follows

$$\left|\frac{1}{l}\sum_{k=1}^{l}\|\mathbf{E}[AX^{(k)}|X^{(k)}_{-I}]\|^2 - \mathbf{E}\left[\|\mathbf{E}[AX|X_{-I}]\|^2\right]\right| = \left|\mathrm{Tr}\left(A\left(\frac{1}{l}\sum_{k=1}^{l}S^{(k)} - \mathbf{E}[S]\right)A^\top\right)\right|$$

Assume momentarily that we somehow know that

$$\left\|\frac{1}{l}\sum_{k=1}^{l}S^{(k)} - \mathbf{E}[S]\right\|_2 \le t \tag{24}$$

Then, by the previous calculation and the definition of $A\in\mathcal{A}_\epsilon$, we would have

$$\begin{aligned}\left|\mathrm{Tr}\left(A\left(\frac{1}{l}\sum_{k=1}^{l}S^{(k)} - \mathbf{E}[S]\right)A^\top\right)\right| &\le \sum_{i=1}^{n}\left|A_i^\top\left(\frac{1}{l}\sum_{k=1}^{l}M^{(k)} - \mathbf{E}[M]\right)A_i\right|\\ &\le t\cdot\sum_{i=1}^{n}\|A_i\|_2^2 = t\cdot\|\hat{J}-J^*\|_F^2\ .\end{aligned}$$

In the above, we have used A_i to denote the i-th row of matrix A, together with the fact that A is symmetric. Thus, to prove the claim, it suffices to establish (24) with high probability. We turn our attention to that task now.

First of all, we notice that $S^{(k)}$ are sampled independently from the same distribution of matrices. Since X is a binary vector, each entry of the random vector $\mathbf{E}[AX|X_{-I_j}]$ lies within $[-1,1]$. Thus, so each $S^{(k)}$ is a symmetric rank-1 matrix with all its entries bounded by 1 in absolute value. Therefore we know that $\|S^{(k)} - \mathbf{E}[S]\|_2 \leq 2n$. And finally, $\|\mathbf{E}[(S^{(k)} - \mathbf{E}[S])^2]\|_2 \leq 4n^2$. Therefore, if we use the matrix Bernstein concentration inequality (see Theorem 1.6.2 in Tropp et al. (2015)), we can have

$$\mathbf{Pr}\left[\left\|\frac{1}{l}\sum_{k=1}^{l} S^{(k)} - \mathbf{E}[S]\right\|_{op} > t\right] = 2n\exp\left(\frac{-lt^2/2}{4n^2 + 2nt/3}\right) \quad .$$

This concludes the proof. ■

The next step will be to use this concentration property to obtain a uniform upper bound for the first derivative of the pseudo-likelihood. We start by proving such a bound for a single direction. Let us define the following event, which depends on the values of $X^{(k)}_{-I_j}$ for $k = 1, \ldots, l$.

$$E_{j,u} := \left\{\left|\frac{1}{l}\sum_{k=1}^{l} \|\mathbf{E}[(\hat{J} - J^*)X^{(k)}|X^{(k)}_{-I_j}]\|^2 - \mathbf{E}\left[\|\mathbf{E}[(\hat{J} - J^*)X|X_{-I_j}]\|^2\right]\right| \leq u \cdot \|A\|_F^2 \ , \forall A\right\}$$

Lemma 32 *Suppose $X^{(1)}, \ldots, X^{(l)} \in \{-1,1\}^n$ are independent samples from an Ising model with interaction matrix J^* satisfying Assumption 29. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with $\|A\|_\infty \leq M$. Suppose we condition on the values of $X^{(k)}_{-I_j}$ for all k and that these values are such that $E_{j,u}$ holds. Then we have that with probability at least*

$$1 - \exp\left(-c\min(t, t^2)\left(l \cdot \left(\mathbf{E}[\|AX\|_2^2]\right)\right)\right),$$

we have that

$$\left|\frac{\partial \phi_j(J^*)}{\partial A}\right| \leq Ct \cdot l \cdot \left(\mathbf{E}[\|AX\|_2^2]\right) \quad .$$

Proof We know that if we choose $\eta = 1/3$ the distribution $X_{I_j}|X_{-I_j}$ satisfy $\|A_{I_j}\|_{op} \leq 1$, satisfy Glauber-MLSI(6) and Po(6). First, we use Theorem 18. Consider a large Ising model, with diagonal matrix blocks A_{I_j}, and also external fields $A_{-I_j,-I_j}X^{(k)}_{-I_j}$. This large Ising model is a tensor product of i.i.d. Ising models of $X_{I_j}|X^{(k)}_{-I_j}$. Therefore, we have, with at least

$$1 - \frac{8}{3}\exp(-C\min(t, t^2)(l \cdot \|A_{I_j}\|_F^2 + \sum_{i=1}^{l} \|\mathbf{E}[A_{I_j}X^{(k)}|X^{(k)}_{i_j}]\|^2))$$

probability, we have

$$\left|\frac{\partial \phi_j(J^*)}{\partial A}\right| \leq t \cdot \left(l \cdot \|A_{I_j}\|_F^2 + \sum_{i=1}^{l} \|\mathbf{E}[A_{I_j}X^{(k)}|X^{(k)}_{i_j}]\|^2\right).$$

If $E_{j,u}$ holds, we have, with at least probability

$$1 - \exp\left(-c\min(t,t^2)\left(l\cdot\left(\|A_{I_j}\|_F^2 + \mathbf{E}[\|\,\mathbf{E}[AX|X_{-I_j}]\|]^2 - u\cdot\|A\|_F^2\right)\right)\right),$$

we have that

$$\left|\frac{\partial\phi_j(J^*)}{\partial A}\right| \le t\cdot l\cdot\left(\|A_{I_j}\|_F^2 + \mathbf{E}[\|\,\mathbf{E}[AX|X_{-I_j}]\|]^2 + u\cdot\|A\|_F^2\right)$$

We now use the following Lemma, which connects the variance of linear functions of Ising models in high temperature with the Frobenius norm and has been repeatedly used in our analysis so far.

Thus, we can write

$$\begin{aligned}
\mathbf{E}[\|AX\|_2^2] &= \sum_{i=1}^n \mathbf{E}[(A_i^\top X)^2] \\
&= \sum_{i=1}^n \mathbf{E}\left[\mathbf{E}[(A_i^\top X)^2|X_{-I_j}]\right] \\
&= \sum_{i=1}^n \left(\mathbf{E}\left[\mathbf{Var}[(A_i^\top X)^2|X_{-I_j}]\right] + \mathbf{E}\left[\left(\mathbf{E}[A_i^\top X|X_{-I_j}]\right)^2\right]\right) \\
&= \sum_{i=1}^n \left(\mathbf{E}\left[\mathbf{Var}[(A_i^\top X)^2|X_{-I_j}]\right]\right) + \mathbf{E}\left[\left\|\mathbf{E}[AX|X_{-I_j}]\right\|_2^2\right]
\end{aligned}$$

Now, we can apply Lemma 22 and Poincaré Inequality to the conditional Ising model conditioned on the values of X_{-I_j}, there are absolute constants c_M, C_M such that

$$c_M\|A_{I_j}\|_F^2 \le \sum_{i=1}^n \left(\mathbf{E}\left[\mathbf{Var}[(A_i^\top X)^2|X_{-I_j}]\right]\right) \le C_M\|A_{I_j}\|_F^2.$$

Thus, the preceding bound implies

$$\mathbf{E}[\|AX\|_2^2] = \Theta\left(\|A_{I_j}\|_F^2 + \|\,\mathbf{E}[AX|X_{-I_j}]\|^2\right)$$

Thus, by adjusting the constants, we know that for any $A\in\mathbb{R}^{n\times n}$, with probability at least

$$1 - \exp\left(-c\min(t,t^2)\left(l\cdot\left(\mathbf{E}[\|AX\|_2^2] - u\cdot\|A\|_F^2\right)\right)\right)$$

we have that

$$\left|\frac{\partial\phi_j(J^*)}{\partial A}\right| \le t\cdot l\cdot\left(\mathbf{E}[\|AX\|_2^2] + u\cdot\|A\|_F^2\right) \quad .$$

Notice also that by Lemma 22, we can absorb the term $\|A\|_F^2$ inside $\mathbf{E}[\|AX\|^2]$ in the upper bound, with the possibility of incurring an extra constant factor. Also, by choosing u to be a small enough constant, again by Lemma 22 we can write

$$\mathbf{E}[\|AX\|_2^2] - u\|A\|_F^2 \ge \frac{1}{2}\,\mathbf{E}[\|AX\|_2^2]$$

Thus, under even $E_{j,u}$ for this choice of constant u, we have that with probability at least

$$1 - \frac{8}{3} \exp\left(-c \min(t, t^2) \left(l \cdot \left(\mathbf{E}[\|AX\|_2^2]\right)\right)\right)$$

we have that

$$\left|\frac{\partial \phi_j(J^*)}{\partial A}\right| \leq Ct \cdot l \cdot \left(\mathbf{E}[\|AX\|_2^2]\right) \quad .$$

■

We would like to prove that the bound of Lemma 32 holds uniformly for all matrices A in a given set. To do that, we first establish a Lipschitzness property of the first derivative of the pseudo-likelihood, similar to the one that was established for the pseudolikelihood itself in Dagan et al. (2021).

Lemma 33 *Let A, B be two symmetric matrices with $\|A\|_\infty, \|B\|_\infty \leq M$. Then*

$$\left|\frac{\partial \phi_j(J^*)}{\partial A} - \frac{\partial \phi_j(J^*)}{\partial B}\right| \leq 2 \cdot l \cdot n \cdot \|A - B\|_2$$

Proof We have that

$$\begin{aligned}
\left|\frac{\partial \phi_j(J^*)}{\partial A} - \frac{\partial \phi_j(J^*)}{\partial B}\right| &= \left|\sum_{k=1}^{l} \sum_{i \in I_j} \left((A_i - B_i)\, X^{(k)}\right) \left(\tanh(J_i^* X^{(k)}) - X_i^{(k)}\right)\right| \\
&\leq 2 \sum_{k=1}^{l} \sum_{i=1}^{n} \left|(A_i - B_i) X^{(k)}\right| \\
&\leq 2\sqrt{n} \sum_{k=1}^{l} \sqrt{\sum_{i=1}^{n} \left|(A_i - B_i) X^{(k)}\right|^2} \\
&\leq 2 \cdot l \cdot n \cdot \|A - B\|_2 \quad .
\end{aligned}$$

In the last step, we used the fact that $\|X^{(k)}\|_2 \leq \sqrt{n}$ for all k and the definition of the operator norm. ■

Our method of bounding the first derivative uniformly is similar to the one employed in Dagan et al. (2021). In particular, we construct a net over the set of matrices and then take a union bound over all the elements of the set to bound the first derivative for all these points. If the radius is chosen small enough, then the upper bound of the first derivative extends to all elements in our set.

Since the probability of failure is governed by $\mathbf{E}[\|AX\|^2]$, we need to choose a set of matrices for which this quantity is large, if we wish to prove high probability bounds. Thus, we define the following set of matrices.

$$\mathcal{R}_s := \{A : \mathbf{E}[\|AX\|^2] \geq s\}$$

Lemma 34 *Suppose $X^{(1)}, \dots, X^{(l)} \in \{-1,1\}^n$ are independent samples from an Ising model with interaction matrix J^* satisfying Assumption 29. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with $\|A\|_\infty \le M$. Consider the net $\mathcal{U}_s := \mathcal{N}(\mathcal{R}_s, \|\cdot\|_2, \theta)$. Then,*

$$\left|\frac{\partial\phi(J^*)}{\partial A}\right| \le \frac{8CM}{\eta}\cdot t\cdot l\cdot \mathbf{E}[\|AX\|_2^2] + 2\cdot l\cdot n\cdot\theta \ \ ,\forall A\in\mathcal{R}_s$$

with probability at least

$$1-\frac{8}{3}(\log n)\cdot|\mathcal{U}_s|\cdot\exp\left(-c\min(t,t^2)\cdot l\cdot s\right) - 2n(\log n)\cdot\exp\left(-\frac{l\cdot u^2/2}{4n^2+2u\cdot n/3}\right)$$

Proof

By taking a union bound over the elements of $\mathcal{U}_s$, by definition of the set $\mathcal{R}_s$ we know that

$$\left|\frac{\partial\phi_j(J^*)}{\partial A}\right| \le C\cdot t\cdot l\cdot\left(\mathbf{E}[\|AX\|_2^2]\right) \ \ ,\forall A\in\mathcal{U}_s$$

with probability at least

$$1-|\mathcal{U}_s|\cdot\exp\left(-c\min(t,t^2)\cdot l\cdot s\right) \ .$$

Using the Lipschitzness of the first derivative, this implies that with the same probability, we have

$$\left|\frac{\partial\phi_j(J^*)}{\partial A}\right| \le C\cdot t\cdot l\cdot\mathbf{E}[\|AX\|_2^2] + 2\cdot l\cdot n\cdot\theta \ \ ,\forall A\in\mathcal{R}_s$$

This bound holds conditional on $X^{(1)}, \dots, X^{(l)}$, assuming they have values that satisfy the event $E_{j,u}$. But we have already bounded the probability that this event occurs in Lemma 31. Thus, for the choice of constant u that we have made, we have established that with probability at least

$$1-|\mathcal{U}_s|\cdot\exp\left(-c\min(t,t^2)\cdot l\cdot s\right) - 2n\exp\left(-\frac{l\cdot u^2/2}{4n^2+2u\cdot n/3}\right),$$

it holds

$$\left|\frac{\partial\phi_j(J^*)}{\partial A}\right| \le C\cdot t\cdot l\cdot\mathbf{E}[\|AX\|_2^2] + 2\cdot l\cdot n\cdot\theta \ \ ,\forall A\in\mathcal{R}_s.$$

We will see in the sequel what the optimal way is to adjust these parameters. Finally, by taking another union bound with respect to all different subsets I_j and using (23), we have that

$$\left|\frac{\partial\phi(J^*)}{\partial A}\right| \le \frac{8CM}{\eta}\cdot t\cdot l\cdot\mathbf{E}[\|AX\|_2^2] + 2\cdot l\cdot n\cdot\theta \ \ ,\forall A\in\mathcal{R}_s,$$

with probability at least

$$1-(\log n)\cdot|\mathcal{U}_s|\cdot\exp\left(-c\min(t,t^2)\cdot l\cdot s\right) - 2n(\log n)\cdot\exp\left(-\frac{l\cdot u^2/2}{4n^2+2u\cdot n/3}\right).$$

■

This is the uniform bound on the first derivative that we were aiming for. We will see how to pick the value of s later.

We now focus on the second derivative. We start with a Lipschitzness property for the second moment in the second derivative.

Lemma 35 *Let A, B be two symmetric matrices with $\|A\|_\infty, \|B\|_\infty \leq M$. Then*

$$\left|\sum_{k=1}^{l}\left(\|AX^{(k)}\|_2^2 - \|BX^{(k)}\|_2^2\right)\right| \leq M \cdot l \cdot n \cdot \|A-B\|_2$$

Proof We have that

$$\begin{aligned}\left|\sum_{k=1}^{l}\left(\|AX^{(k)}\|_2^2 - \|BX^{(k)}\|_2^2\right)\right| &= \left|\sum_{k=1}^{l}\sum_{i=1}^{n}\left((A_iX^{(k)})^2 - (B_iX^{(k)})^2\right)\right| \\ &\leq \sum_{k=1}^{l}\sum_{i=1}^{n}|(A_i-B_i)X^{(k)}| \cdot |(A_i+B_i)X^{(k)}| \\ &\leq \sum_{k=1}^{l}\sqrt{\sum_{i=1}^{n}|(A_i-B_i)X^{(k)}|^2} \cdot \sqrt{\sum_{i=1}^{n}|(A_i+B_i)X^{(k)}|^2} \\ &\leq M \cdot l \cdot n \cdot \|A-B\|_2\end{aligned}$$

■

We are now ready to state the uniform guarantee for the second derivative. The proof will again be based on Lemma 31 to avoid unnecessary union bounds.

Lemma 36 *Suppose $X^{(1)}, \ldots, X^{(l)} \in \{-1,1\}^n$ are independent samples from an Ising model with interaction matrix J^* satisfying Assumption 29. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with $\|A\|_\infty \leq M$. Consider the net $\mathcal{U}_s := \mathcal{N}(\mathcal{R}_s, \|\cdot\|_2, \theta)$. Then, with probability at least*

$$1 - |\mathcal{U}_s|\exp\left(-c \cdot l \cdot s\right) - 2n \cdot \exp\left(-\frac{l \cdot u^2/2}{4n^2 + 2u \cdot n/3}\right)$$

we have

$$\frac{\partial^2\phi(J)}{\partial A^2} \geq C \cdot l \cdot \mathbf{E}[\|AX\|^2] - M \cdot l \cdot n \cdot \theta \ \ ,\forall A \in \mathcal{R}_s$$

Proof

We use the same argument as the first derivative. If we take $\eta = 1/3$, take any j, the distribution of each $X_{I_j}|X_{-I_j}$ is Glauber MLSI(6) and Po(6). When we do the tensor product of all the samples, the same Glauber MLSI and Poincaré inequality holds. By Lemma 23, we have that there exists absolute constants c, C such that

$$\frac{\partial^2\phi(J)}{\partial A^2} \geq C\left(l\|A_{I_j}\|_F^2 + \sum_{k=1}^{l}\|\,\mathbf{E}[AX^{(k)}|X^{(k)}_{-I_j}]\|^2\right)$$

with probability at least

$$1 - \exp\left(-c \cdot \left(l \cdot \|A_{I_j}\|_F^2 + \sum_{k=1}^{l}\|\,\mathbf{E}[AX^{(k)}|X^{(k)}_{-I_j}]\|^2\right)\right)$$

The previous arguments have already established that

$$\|A_{I_j}\|_F^2 + \|\mathbf{E}[AX|X_{-I_j}]\|_2^2 = \Theta(\mathbf{E}[\|AX\|^2]).$$

Using the concentration of Lemma 31 as before, we can establish that if event $E_{j,u}$ holds for a small enough constant u, then conditional on $X_{-I_j}^{(k)}$ for $k = 1, \ldots, l$ we have, there exists absolute constants c, C such that

$$\frac{\partial^2 \phi(J)}{\partial A^2} \geq C \cdot l \cdot \mathbf{E}[\|AX\|^2]$$

with probability at least

$$1 - \exp\left(-c \cdot l \cdot \mathbf{E}[\|AX\|^2]\right)$$

We can now implement the exact same union bound argument that we had for the first derivative. The only thing we need to check is the Lipschitzness of the second derivative, which will determine the size of our net.

Now, again by considering the net $\mathcal{U}_s := \mathcal{N}(\mathcal{R}_s, \|\cdot\|_2, \theta)$, taking a union bound over its elements and using the lipschitzness of the second derivative, in exactly the same fashion as with the first derivative, we get that conditioned on $E_{j,u}$

$$\frac{\partial^2 \phi(J)}{\partial A^2} \geq C \cdot l \cdot \mathbf{E}[\|AX\|^2] - M \cdot l \cdot n \cdot \theta \ \ , \forall A \in \mathcal{R}_s \tag{25}$$

with probability at least

$$1 - |\mathcal{U}_s| \exp\left(-c \cdot l \cdot s\right)$$

Removing the conditioning and using Lemma 31, we have that (25) holds with probability at least

$$1 - |\mathcal{U}_s| \exp\left(-c \cdot l \cdot s\right) - 2n \cdot \exp\left(-\frac{l \cdot u^2/2}{4n^2 + 2u \cdot n/3}\right)$$

■

We are now ready to use the above lemmas to argue about the value of the pseudolikelihood for matrices that are "far" from J^*.

Theorem 37 *Let $X_1, \ldots, X_l$ be independent samples drawn from an Ising model with interaction matrix J^* satisfying Assumption 29. Let $\hat{J} \in \mathbb{R}^{n \times n}$ be the estimate of J^* that is obtained by maximizing the pseudo-likelihood function (6), i.e.*

$$\hat{J}^{(l)} := \operatorname*{argmax}_{J^{(l)} \in \mathcal{R}^{(l)}} \phi(J; X) \ \ ,$$

where $\mathcal{R} = \{J \in \mathcal{S}_0^n : \|J\|_\infty \leq M$. Then, for any $\epsilon \in (0,1)$ and $\delta > 0$, if

$$l = \tilde{\Omega}\left(\frac{n^2\left(\log(1/(\delta\varepsilon)) + \log n\right)}{\epsilon}\right)$$

then with probability at least $1 - \delta$

$$\mathop{\mathbf{E}}_{J^*}[\|(\hat{J} - J^*)X\|_2^2] \leq \epsilon \ \ ,$$

where $\tilde{O}$ hides $\exp(M)$ factors.

Proof For any matrix $J \in \mathcal{J}_M$, an application of Taylor's Theorem yields

$$\phi(J) = \phi(J^*) + \frac{\partial \phi(J^*)}{\partial A}|_{A=J-J^*} + \frac{1}{2}\frac{\partial^2 \phi(J_\xi)}{\partial A^2}|_{A=J-J^*}$$

where J_ξ belongs in the line segment connecting J, J^*. Thus, using the preceding arguments, we know that with probability at least

$$1 - (\log n) \cdot |\mathcal{U}_s| \cdot \exp\left(-c\min(t, t^2) \cdot l \cdot s\right) - 2n(\log n) \cdot \exp\left(-\frac{l \cdot u^2/2}{4n^2 + 2u \cdot n/3}\right),$$

we have

$$\phi(J) \geq \phi(J^*) + C \cdot l \cdot \mathbf{E}[\|(J-J^*)X\|_2^2] - C' \cdot t \cdot l \cdot \mathbf{E}[\|(J-J^*)X\|_2^2] - C'' \cdot l \cdot n \cdot \theta \ \ , \forall J : J - J^* \in \mathcal{R}_s. \tag{26}$$

We now show how to choose the various parameters. First, we choose t to be a small enough constant so that $C't < C/4$. We also choose $\theta = O(s/n)$ so that $C'' \cdot l \cdot n \cdot \theta \leq Cs/4$. These choices mean that (26) can be written as

$$\phi(J) \geq \phi(J^*) + \frac{C}{4} \cdot l \cdot s \ \ , \forall J : J - J^* \in \mathcal{R}_s \tag{27}$$

This means that all matrices J such that $J - J^* \in \mathcal{U}_s$ have a higher negative pseudolikelihood value than J^*, which means that they will not be selected by the optimization procedure. Thus, this allows us to conclude that $\hat{J} - J^* \notin \mathcal{R}_s$, which implies that $\mathbf{E}[\|(\hat{J} - J^*)X\|_2^2] \leq s$. We now turn to analyze the probability that this event occurs. Let us set $s = \epsilon$. We would like to have

$$(\log n) \cdot |\mathcal{U}_s| \cdot \exp\left(-c\min(t, t^2) \cdot l \cdot \epsilon\right) \leq \frac{\delta}{2} \tag{28}$$

for the specified error probability $\delta > 0$. As we argued in the previous section, using Corollary 4.1.15 from Artstein-Avidan et al. (2021) gives

$$|\mathcal{U}_s| \leq \left(M + \frac{2}{\theta}\right)^{n^2} = O\left(\left(M + \frac{2n}{\epsilon}\right)^{n^2}\right)$$

Thus, if we choose

$$l \geq C\frac{n^2 \log(n/\varepsilon) + \log\log n + \log(1/\delta)}{\epsilon} \tag{29}$$

for some constant $C > 0$ that depends on M, we can satisfy (28). We would also like to have

$$2n(\log n) \cdot \exp\left(-\frac{l \cdot u^2/2}{4n^2 + 2u \cdot n/3}\right) \leq \frac{\delta}{2} \tag{30}$$

In the above, u is a sufficiently small constant. To satisfy (30), it suffices to choose

$$l \geq C\left(n^2 \log n + n^2 \log(1/\delta)\right) \tag{31}$$

samples. The conditions (29) and (31) give us the final sample complexity.

■

C.1. Proof of Proof of Corollary 10

Finally, we give a formal version of Corrolary 10 and then its proof.

Corollary 38 *Suppose we are in the setting of Theorem 37. Additionally, assume there exist constants $\gamma, C > 0$, such that $\mathbf{Pr}_{J^*}$ satisfies $(\gamma/n, C)$-regularity regularity. Then, for any $\epsilon > 0$, if $l = \tilde{\Omega}(n^3/\epsilon^2)$, with probability $1 - o(1)$ over the choice of samples $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \leq \epsilon$.*

Proof Using the same representation for KL between Ising models as Lemma 20 of Dagan et al. (2021), we get that for some matrix $J_\xi = \xi J^* + (1-\xi)\hat{J}$

$$KL\left(P_{\hat{J}} \middle\| P_{J^*}\right) = \frac{1}{2} \operatorname*{\mathbf{Var}}_{X \sim J_\xi}\left[X^\top \left(J^* - \hat{J}\right) X\right] \ ,$$

where in the above, the notation X is sampled from an Ising model with interaction matrix J_ξ. The Cauchy-Schwarz inequality now implies that

$$\begin{aligned}
\operatorname*{\mathbf{Var}}_{X \sim J_\xi}\left[X^\top \left(J^* - \hat{J}\right) X\right] &\leq \operatorname*{\mathbf{E}}_{J_\xi}[(X^\top \left(J^* - \hat{J}\right) X)^2] \\
&\leq n \cdot \operatorname*{\mathbf{E}}_{J_\xi}[\|(\hat{J} - J^*)X\|_2^2] \\
&= n \cdot \frac{\mathbf{E}_{J_\xi}[\|(J_\xi - J^*)X\|_2^2]}{(1-\xi)^2}
\end{aligned}$$

Without loss of generality, assume that $\epsilon^2 < \gamma$. By Theorem 37, we know that if $l = \tilde{\Omega}(n^3/\epsilon^2)$, we have with probability $1 - o(1)$

$$\operatorname*{\mathbf{E}}_{J^*}[\|(\hat{J} - J^*)X\|_2^2] \leq \frac{\epsilon^2}{n}$$

Then, by scaling, this implies

$$\operatorname*{\mathbf{E}}_{J^*}[\|(J_\xi - J^*)X\|_2^2] \leq \frac{\epsilon^2}{n} \cdot (1-\xi)^2 \leq \frac{\epsilon^2}{n}$$

Thus, since we know J^* satisfies $(\gamma/\sqrt{n}, C)$-regularity, we have that

$$\operatorname*{\mathbf{E}}_{J_\xi}[\|(J_\xi - J^*)X\|_2^2] \leq C\frac{\epsilon^2 \cdot (1-\xi)^2}{n}$$

which implies

$$\mathrm{KL}\left(P_{\hat{J}} \middle\| P_{J^*}\right) \leq C \cdot \epsilon^2$$

Using Pinsker's inequality as in Lemma 40 concludes the proof. ■

Appendix D. Applications of Learning with MLSI

In this Section, we present the proofs for the applications of Theorem 28. The step that remains is to bound the total variation distance between two Ising models by the Frobenius norm of the difference of their interaction matrices. We give two such bounds. The first is cruder and makes no additional assumptions about the matrices.

Lemma 39 *Suppose μ, μ^* are the distributions corresponding to two Ising models with interaction matrices $J, J^* \in \mathcal{S}_0^n$ and zero external fields. Then, we have the following property:*

$$\|\mu - \mu^*\|_{TV} \leq n\|J - J^*\|_F.$$

We give our version of the proof in Section E. Note that an alternative proof can be given using the technique in Klivans and Meka (2017) Lemma 7.3.

The second Lemma additionally assumes that both matrices, as long as any matrix in the line that contains them, satisfy MLSI. In that case, we can obtain much more precise guarantees without losing polynomial factors.

Lemma 40 *Suppose $J_1, J_2 \in \mathcal{S}_0^n$ are such that for every $t \in [0,1]$, $\mathbf{Pr}_{J_t}$ satisfies MLSI(ρ), where $J_t = tJ_1 + (1-t)J_2$. Then,*

$$\mathrm{TV}(P_{J_1}, P_{J_2}) \leq \rho \cdot \|J_1 - J_2\|_F$$

Proof First, using Pinsker's inequality, we get

$$TV(P_{J_1}, P_{J_2}) \leq \sqrt{\frac{KL\left(P_{J_1} \| P_{J_2}\right)}{2}} \quad .$$

Thus, it suffices to bound $KL\left(P_{J_1} \| P_{J_2}\right)$. For this, we rely on a standard calculation for exponential families that connects the KL divergence to the variance of the sufficient statistic. In particular, following the derivation in Lemma 20 of Dagan et al. (2021), we get that for some matrix J_ξ contained in the line segment connecting J_1 and J_2

$$KL\left(P_{J_1} \| P_{J_2}\right) = \frac{1}{2} \operatorname*{\mathbf{Var}}_{X \sim J_\xi}\left[X^\top \left(J_1 - J_2\right) X\right] \quad ,$$

where in the above, the notation X is sampled from an Ising model with interaction matrix J_ξ. Since J_ξ satisfy MLSI(ρ), it also satisfies Po(2ρ)), which yields

$$\begin{aligned}
\operatorname*{\mathbf{Var}}_{X \sim S}\left[X^\top \left(J_1 - J_2\right) X\right] &\leq 2\rho \operatorname*{\mathbf{E}}_{X \sim S}\left[\|(J_1 - J_2)X\|_2^2\right] \\
&= 2\rho \sum_{i=1}^{n} \operatorname*{\mathbf{Var}}_{X \sim S}\left[(J_1 - J_2)_i^\top X\right] \\
&\leq 4\rho^2 \sum_{i=1}^{n} \|(J_1 - J_2)_i\|_2^2 \\
&= 4\rho^2 \|J_1 - J_2\|_F^2
\end{aligned}$$

This concludes the proof. ■

We are now ready to present the proof of the applications.

D.1. Application: SK/diluted SK model

First, using Lemma 39 and Theorem 28, we can derive the following corollary:

Corollary 41 *Let $X_1, \dots, X_l$ be independent samples drawn from an Ising model with interaction matrix J^* satisfying Assumption 17. Let $\hat{J} \in \mathbb{R}^{n\times n}$ be the estimate of J^* that is obtained by maximizing the pseudo-likelihood function (6) for a set of matrices $\mathcal{R} \subseteq \mathcal{S}_0^n$, which is convex and admits efficient projections in Frobenius norm. Then, for any $\varepsilon \in (0,1)$, if $l = \tilde{O}\left(\frac{n^4 \log(1/(\delta\varepsilon))}{\varepsilon^2}\right)$ then with probability at least $1-\delta$, $TV(P_{\hat{J}}, P_{J^*}) \le \varepsilon$, where $\tilde{O}$ hides sub-polynomial factors of n, and other terms involving λ, ρ and $h_{\max}$. Moreover, we can implement MPLE in polynomial time.*

Proof The proof follows immediately by setting $\epsilon = \epsilon'/n$ in the guarantees of Theorem 28. The optimal value of the pseudolikelihood can be found in polynomial time using projected gradient descent. ■

Corollary 41 says that if we can find a set $\mathcal{R}$ that is convex and admits efficient projections, we can solve the pseudolikelihood estimation problem in polynomial time. We instantiate this fact in the cases of SK/diluted SK model now.

Corollary 42 *Suppose we are given l independent samples $X^{(1)}, \dots, X^{(l)} \sim \mathbf{Pr}_{J^*}$, where J^* is sampled according to the SK-model with $\beta < C$, where $C \approx 0.295$. That is, every $J_{ij} = J_{ji}$ is chosen i.i.d. from $\mathcal{N}(0, \beta^2/n)$. Then, there is a polynomial time algorithm (MPLE) that produces an estimate $\hat{J} \in \mathcal{S}_0^n$, such that with probability $1 - o(1)$ over the choice of samples and the choice of matrix J^* we have $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \le \epsilon$, as long as $l = \tilde{\Omega}(n^4/\epsilon^2)$.*

Proof By Anari et al. (2024b), when $\beta < C$, with $1 - o(1)$ probability over the choice of random matrix J^* it satisfies ATE, which by Chen et al. (2021) means it also satisfies MLSI. Therefore, we can run MPLE on the set of matrices $\mathcal{R} = \{J \in \mathcal{S}_0^n : \|J\|_{op} \le 4\}$. We know that $J^* \in \mathcal{R}$ with probability $1 - o(1)$. Also, it is clear we can efficiently project to $\mathcal{R}$ by eigenvalue clipping. Thus, applying Corollary 41 finishes the proof. ■

Similar guarantees can be obtained for the diluted SK model.

Corollary 43 *Suppose we are given l independent samples $X^{(1)}, \dots, X^{(l)} \sim \mathbf{Pr}_{J^*}$, where J^* is sampled according to the diluted SK-model with $\beta < C$, where $C \approx 0.295$. That is, consider a random d-regular graph G. If i and j are not connected, $J_{ij} = 0$. If i and j are connected, then $J_{ij} = J_{ji}$ is chosen from $\{\frac{\beta}{\sqrt{d-1}}, -\frac{\beta}{\sqrt{d-1}}\}$ with the same probability, independently for all such pairs. Then, there is a polynomial-time algorithm (MPLE) that produces an estimate $\hat{J} \in \mathcal{S}_0^n$, such that with probability $1 - o(1)$ over the choice of samples and the choice of matrix J^* we have $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \le \epsilon$, as long as $l = \tilde{\Omega}(n^4/\epsilon^2)$.*

Proof The proof is similar to Corollary 42. By Anari et al. (2024b), when $\beta < C$, with $1 - o(1)$ probability it satisfies ATE, thus also MLSI with a constant that depends on the distance of β from C. Thus, we can use Corollary 41 to obtain the result. ■

D.2. Application: Spectrally Bounded Models

For spectrally bounded models, the set $\mathcal{R}$ can be naturally restricted to only include matrices that satisfy MLSI. We thus get optimal guarantees for TV learning.

Corollary 44 *Let $X_1, \ldots, X_l$ be independent samples drawn from an Ising model with interaction matrix $J^* \in \mathcal{S}_0^n$ that belongs in the set $\mathcal{R} = \{J \in \mathcal{S}_0^n : \lambda_{max}(J) - \lambda_{min}(J) \leq 1 - \alpha\}$, for some $\alpha \in (0,1)$, and zero external field. Let $\hat{J} \in \mathbb{R}^{n \times n}$ be the estimate of J^* that is obtained by maximizing the pseudo-likelihood function (6) over $\mathcal{R}^{(l)}$, given l independent samples from J^*. Then, for any $\varepsilon < 1$, with probability at least $1 - \delta$*

$$TV(P_{\hat{J}}, P_{J^*}) \leq \varepsilon \ ,$$

whenever

$$l \geq \exp\left(C_0\left(\sqrt{\log(n/\delta)\cdot\alpha^{-1}} + \alpha^{-5/4}\varepsilon\right)\right)\cdot\frac{n^2\log(1/(\delta\varepsilon)}{\varepsilon^2}.$$

Where C_0 is a universal constant. Moreover, the algorithm can be implemented in polynomial time.

Proof First, let us argue that $\mathcal{R}$ is a convex set. We know that for two symmetric matrices, using the Rayleigh quotient, we can derive that for any $t \in (0,1)$

$$\lambda_{\max}(tA_1 + (1-t)A_2) \leq t\lambda_{\max}(A_1) + (1-t)\lambda_{\max}(A_2)$$
$$\lambda_{\min}(tA_1 + (1-t)A_2) \geq t\lambda_{\min}(A_1) + (1-t)\lambda_{\min}(A_2)$$

so if $A_1, A_2 \in \mathcal{R}$, $tA_1 + (1-t)A_2 \in \mathcal{R}$.

By Lemma 12, every $J \in \mathcal{R}$ satisfies MLSI($1/\alpha$). Also, clearly $\|J\|_{op} < 1$ for every $J \in \mathcal{R}$, since the zero diagonal implies J will have both positive and negative eigenvalues. Thus, combining Theorem 28 and Lemma 40 gives the desired guarantee in TV.

Now, let us argue about the computational efficiency of the method. Since the pseudolikelihood function is convex, to find $\hat{J}$ we can use the projected gradient descent algorithm as in Theorem 3.2 of Bubeck et al. (2015). Thus, we only need to argue that at every step we can efficiently project on the set $\mathcal{R}$.

We have shown that $\mathcal{R}$ is a convex set, and the distance from it, which is measured in Frobenius norm, is a convex function. Thus, one way of computing the projection would be using the result in Lee et al. (2018), where they optimize the distance in polynomial time using only a membership oracle for $\mathcal{R}$, which is, of course, easy to implement in our case. Alternatively, one could write the projection as a semi-definite program (SDP) Boyd and Vandenberghe (2004) and then solve it efficiently using Jiang et al. (2020). For convenience, we spell out the details of the SDP approach below.

The SDP in standard form minimizes $\langle C, X\rangle$ subject to $\langle A_i, X\rangle \leq b_i$ and $X \succeq 0$. We can use a diagonal block matrix X so that $X \succeq 0$ is equivalent to $X_1, X_2, \ldots, X_k \succeq 0$, and furthermore, using the linear constraints $\langle A_i, X\rangle \leq b_i$ (both $\langle A, X\rangle \leq b$ and $\langle -A, X\rangle \leq b$ to build an equation) we can encode linear relationships between matrices $X_1, X_2, \ldots, X_k$. Therefore, we can actually write several, rather than one, positive semi-definite constraints of matrices, where each element has a linear relation to the other.

Therefore, we can first encode two matrices J (the matrix that needs to be projected) and J' (the projected matrix), and three values λ_1, λ_2, t in X_1's off-diagonal and put free variables on the

diagonal of X_1 so that $X_1 \succeq 0$ if we can make the diagonal large enough. In the linear constraints, we encode the information that J and J' are both zero-diagonal symmetric matrices. Then, we encode $\lambda_1 I - J' \succeq 0$ and $J' - \lambda_2 I \succeq 0$ and $\lambda_1 - \lambda_2 \leq 1 - \alpha$ to make the constraint $\lambda_{\max}(J') - \lambda_{\min}(J') \leq 1 - \alpha$. Finally, we flatten the elements in $J - J'$ into a column vector v, and we add the constraint $\begin{pmatrix} I & v \\ v^\top & t \end{pmatrix} \succeq 0$ which is equivalent to $t \geq \|J - J'\|_F^2$. Finally, we optimize t to make it as small as possible. We solve this SDP to find J', which is the projection from J to $\mathcal{R}$. This SDP has a polynomial size in n and a polynomial number of constraints. By Jiang et al. (2020) (the main result), we can get an efficient algorithm solving the SDP, thus we can project on $\mathcal{R}$ efficiently. ■

D.3. Application: Antiferromagnetic Expanders

Corollary 45 *Given $0 < \alpha < 1$ and $c > 0$. Let $\mathbf{1}$ be the all-one vector. Let $\mathcal{A}$ be the set of matrices such that for all $A \in \mathcal{A}$, $A\mathbf{1} = 0$ and $0 \preceq A \preceq (1-\alpha)I$. Suppose we are given l independent samples $X^{(1)}, \ldots, X^{(l)} \sim \mathbf{Pr}_{J^*}$, where J^* is from*

$$\mathcal{R} = \mathcal{S}_0^n \cap \{\exists r \in \mathbb{R}, 0 \leq t \leq c, A \in \mathcal{A}, \text{ s.t. } J^* + rI = -\frac{t}{n}\mathbf{1}\mathbf{1}^\top + A\}.$$

Then, there is a polynomial time algorithm (MPLE) that produces an estimate $\hat{J} \in \mathcal{R}$, such that with high probability over the choice of samples and the choice of matrix J^* we have $\mathrm{TV}(\mathbf{Pr}_{\hat{J}}, \mathbf{Pr}_{J^*}) \leq \epsilon$, as long as $l = \tilde{\Omega}(n^2/\epsilon^2)$.

Proof First, we know that, from Anari et al. (2024b), any matrix in $\mathcal{R}$ satisfies ATE, and thus satisfies modified-LSI. Following the previous recipe, we just need to prove that $\mathcal{R}$ is convex. First, we know that $\mathcal{A}$ is convex. Since the average of two positive semi-definite (PSD) matrices is PSD, for any $A_1, A_2 \in \mathcal{A}$, $(A_1 + A_2)/2 \succeq 0$, and $(1-\alpha)I - (A_1 + A_2)/2 \succeq 0$. Therefore, $(A_1 + A_2)/2 \in \mathcal{A}$. Therefore, for $J_1, J_2 \in \mathcal{R}$, there exists real numbers $r_1, r_2, t_1, t_2, A_1, A_2$, such that $J_i + r_i I = \frac{t_i}{n}\mathbf{1}\mathbf{1}^\top + A_i$ for $i = 1, 2$. Therefore, we have $(J_1 + J_2)/2 + \frac{r_1 + r_2}{2} I = \frac{(t_1+t_2)/2}{n}\mathbf{1}\mathbf{1}^\top + (A_1 + A_2)/2$, and thus $(J_1 + J_2)/2 \in \mathcal{R}$, and thus $\mathcal{R}$ is convex.

Again, we have shown that $\mathcal{R}$ is a convex set, and we are optimizing a convex function. And the constraints of being a matrix in $\mathcal{R}$ are spectral or linear, as in the proof of Corollary 44. Thus, we can formulate the problem as an SDP again and solve it efficiently Jiang et al. (2020). As in the proof of Corollary 44, we encode matrices J (matrix need to be projected) and J' (the projected matrix) and scalars $\lambda_1, \lambda_2, t/n, v$. We also encode $M = J' - t/n \cdot \mathbf{1}\mathbf{1}^\top$ and $0 \leq t \leq c$ in the linear constraint. We make $M = M^\top$, M has equal-diagonal, and $M\mathbf{1} = 0$ in the linear constraints as well. In addition, we have $\lambda_1 I - M \succeq 0, M - \lambda_2 I \succeq 0$ as a positive semi-definite constraint and $\lambda_1 - \lambda_2 \leq 1 - \alpha$ as a linear constraint. Finally, we repeat the steps in Corollary 44 for the final step of the Frobenius norm. As we can write the SDP like that, we can get an efficient projection algorithm. ■

Remark 46 *We should mention that in this antiferromagnetic expanders model, despite the constraints, it could still be that $\|J^*\|_\infty$ is unbounded. Indeed, consider a Paley graph G of a $4k+1$ type prime p (that is, i and j are connected if and only if $i - j$ is a quadratic residue), and consider $J - I - 2G$, where J is the all-one matrix. We know that G has all the eigenvalues 0 once with the all-one vector, and $\pm\sqrt{p}$ appears $(p-1)/2$ times. Then, we know that if we consider $J^* = \frac{1-\alpha}{2\sqrt{p}}(J - I - 2G) - t \cdot (J - I)$ for $0 < t < c - 1$, it will be in $\mathcal{R}$.*

Appendix E. Bounding the TV distance by the Frobenius norm (Proof of Lemma 39)

Suppose we have two Ising models with interaction matrices $J, J^* \in \mathcal{S}_0^n$ respectively and zero external fields. The following lemma bounds their TV distance in terms of $\|J - J^*\|_F$.

Lemma 47 *Suppose μ, μ^* are the distributions corresponding to two Ising models with interaction matrices $J, J^* \in \mathcal{S}_0^n$ and zero external fields. Then, we have the following property:*

$$\|\mu - \mu^*\|_{TV} \le n\|J - J^*\|_F.$$

Proof We consider the TV distance of the X_i conditioning on μ, μ^*. By Pinsker's inequality, we have

$$2\|\mu - \mu^*\|_{TV}^2 \le KL(\mu\|\mu^*)$$

Consider the symmetric KL distance, we have

$$4\|\mu - \mu^*\|_{TV}^2 \le KL(\mu\|\mu^*) + KL(\mu^*\|\mu) = \mathop{\mathbf{E}}_{X\sim\mu}(X^\top(J - J^*)X) + \mathop{\mathbf{E}}_{X\sim\mu^*}(X^\top(J^* - J)X)$$

We expand one of them, and the second one is analogous.
We can write

$$\mathop{\mathbf{E}}_{X\sim\mu}(X^\top(J - J^*)X) = 2\sum_{i<j}(J_{ij} - J^*_{ij})\mathop{\mathbf{E}}_{X\sim\mu}(X_iX_j) = 2\sum_{i<j}(J_{ij} - J^*_{ij})(2\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1) - 1).$$

Therefore, after we group them, we have

$$\mathop{\mathbf{E}}_{X\sim\mu}(X^\top(J-J^*)X)+\mathop{\mathbf{E}}_{X\sim\mu^*}(X^\top(J^*-J)X) = 4\sum_{i<j}(J_{ij}-J^*_{ij})(\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1)-\mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1))$$

The probability of $X_iX_j = 1$ can be upper-bounded by coupling. First, we consider the distribution of X and X' according to μ and μ^*, respectively. We couple X and X' to achieve the probability $X_{-i-j} \neq X'_{-i-j}$ as small as possible. We call the event $X_{-i-j} \neq X'_{-i-j}$ to be E_1, and $X_{-i-j} = X'_{-i-j}$ to be E_2. Therefore, we can split the difference in probability to be

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X\sim\mu}(X^\top(J - J^*)X) + \mathop{\mathbf{E}}_{X\sim\mu^*}(X^\top(J^* - J)X)\\
=&4\sum_{i<j}(J_{ij} - J^*_{ij})(\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1))\\
\le&4\sum_{i<j}|J_{ij} - J^*_{ij}| \cdot |\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1)|\\
\le&4\sum_{i<j}|J_{ij} - J^*_{ij}| \cdot \left(\mathbf{Pr}(E_1) \cdot \left|\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1|E_1) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1|E_1)\right|\right.\\
&\qquad\left. + \mathbf{Pr}(E_2) \cdot \left|\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1|E_2) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1|E_2)\right|\right)
\end{aligned}$$

For E_1 case, We have $\mathbf{Pr}(E_1) \le \|\mu - \mu^*\|_{TV}$ along with the naive bound

$$|\mathbf{Pr}(X_iX_j = 1|E_1) - \mathbf{Pr}(X_iX_j = 1|E_1)| \le 1.$$

For the E_2 case, we bound $\mathbf{Pr}(E_2) \le 1$, and we upper bound the probability difference by the largest possible difference of the conditional probability. That is:

$$\left|\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1|E_2) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1|E_2)\right| \le \max_{X_{-i-j}} \left|\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1|X_{-i-j}) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1|X_{-i-j})\right|.$$

To sum up, we have the following inequality:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X\sim\mu}(X^\top(J - J^*)X) + \mathop{\mathbf{E}}_{X\sim\mu^*}(X^\top(J^* - J)X)\\
\le& 4\sum_{i<j}|J_{ij} - J^*_{ij}| \cdot \left(\mathbf{Pr}(E_1) \cdot \left|\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1|E_1) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1|E_1)\right|\right.\\
&\qquad\left. + \mathbf{Pr}(E_2) \cdot \left|\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1|E_2) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1|E_2)\right|\right)\\
\le& 4\sum_{i<j}|J_{ij} - J^*_{ij}| \cdot \left(\|\mu - \mu^*\|_{TV} + \max_{X_{-i-j}} \left|\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1|X_{-i-j}) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1|X_{-i-j})\right|\right)
\end{aligned}$$

For Ising model μ with interaction matrix J, denote $h_i = \sum_{k\neq i,j} J_{ik}$, and $h_j = \sum_{k\neq i,j} J_{jk}$ we can calculate that

$$\begin{aligned}
&\mathbf{Pr}(X_iX_j = 1)\\
=&\frac{\exp(J_{ij} + h_i + h_j) + \exp(J_{ij} - h_i - h_j)}{\exp(J_{ij} + h_i + h_j) + \exp(J_{ij} - h_i - h_j) + \exp(-J_{ij} + h_i - h_j) + \exp(-J_{ij} - h_i + h_j)}.
\end{aligned}$$

Denote $v = h_i + h_j$ and $w = h_i - h_j$. So, we have the following

$$\mathbf{Pr}(X_iX_j = 1) = \frac{\exp(J_{ij})\cosh(v)}{\exp(J_{ij})\cosh(v) + \exp(-J_{ij})\cosh(w)} =: F(J_{ij}, u, w).$$

Denote $v^* = \sum_{k\neq i,j} J^*_{ik} + \sum_{k\neq i,j} J^*_{jk}$ and $w^* = \sum_{k\neq i,j} J^*_{ik} - \sum_{k\neq i,j} J^*_{jk}$. Our target is to upper bound $|F(J_{ij}, u, w) - F(J^*_{ij}, u^*, w^*)|$. We take the derivative to show the Lipschitzness of F to achieve this. By taking the derivative directly, we have

$$\begin{aligned}
\frac{\partial F}{\partial J_{ij}} &= \frac{2\cosh(v)\cosh(w)}{(\exp(J_{ij})\cosh(v) + \exp(-J_{ij})\cosh(w))^2}.\\
\frac{\partial F}{\partial v} &= \frac{\sinh(v)\cosh(w)}{(\exp(J_{ij})\cosh(v) + \exp(-J_{ij})\cosh(w))^2}.\\
\frac{\partial F}{\partial v} &= -\frac{\sinh(w)\cosh(v)}{(\exp(J_{ij})\cosh(v) + \exp(-J_{ij})\cosh(w))^2}.
\end{aligned}$$

By AM-GM ineqiality, we have $|\frac{\partial F}{\partial J_{ij}}| \le \frac{1}{2}$. By furthermore $|\sinh(x)| \le |\cosh(x)|$, we have both $|\frac{\partial F}{\partial v}|$ and $|\frac{\partial F}{\partial w}| \le \frac{1}{4}$. Therefore, we have the bound:

$$|F(J_{ij}, u, w) - F(J^*_{ij}, u^*, w^*)| \le \frac{1}{2}|J_{ij} - J^*_{ij}| + \frac{1}{4}|v - v^*| + \frac{1}{4}|w - w^*|$$

Plugging back the expression of v, w, v^*, w^*, we use the triangle inequalty, we have

$$\begin{aligned}|v - v^*| &= \left|\sum_{k\neq i,j} J_{ik} + \sum_{k\neq i,j} J_{jk} - \sum_{k\neq i,j} J^*_{ik} - \sum_{k\neq i,j} J^*_{jk}\right| \\ &= \left|\sum_{k\neq i,j} J_{ik} - J_{ik} + J_{jk} - J^*_{jk}\right| \leq \sum_{k\neq i,j} |J_{ik} - J_{ik}| + |J_{jk} - J^*_{jk}|.\end{aligned}$$

And, we can get the same for $w - w^*$. Therefore, we can deduce that

$$|F(J_{ij}, u, w) - F(J^*_{ij}, u^*, w^*)| \leq \frac{1}{2}|J_{ij} - J^*_{ij}| + \frac{1}{2}\sum_{k\neq i,j} |J_{ik} - J^*_{ik}| + |J_{ik} - J^*_{jk}|.$$

Therefore, we can have the following inequality:

$$\begin{aligned}&\mathop{\mathbf{E}}_{X\sim\mu}(X^\top(J - J^*)X) + \mathop{\mathbf{E}}_{X\sim\mu^*}(X^\top(J^* - J)X) \\ \leq& 4\sum_{i<j} |J_{ij} - J^*_{ij}| \cdot \left(\|\mu - \mu^*\|_{TV} + \max_{X_{-i-j}} \left|\mathop{\mathbf{Pr}}_{X\sim\mu}(X_iX_j = 1|X_{-i-j}) - \mathop{\mathbf{Pr}}_{X\sim\mu^*}(X_iX_j = 1|X_{-i-j})\right|\right) \\ \leq& 4\sum_{i<j} |J_{ij} - J^*_{ij}| \cdot \left(\|\mu - \mu^*\|_{TV} + \frac{1}{2}|J_{ij} - J^*_{ij}| + \frac{1}{2}\sum_{k\neq i,j} |J_{ik} - J^*_{ik}| + |J_{ik} - J^*_{jk}|\right)\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\sum_{i<j} |J_{ij} - J^*_{ij}| = \frac{1}{2}\sum_{i\neq j} |J_{ij} - J^*_{ij}| \leq \frac{n}{2}\|J - J^*\|_F.$$

Also, the rest of that is

$$\begin{aligned}&\sum_{i<j} |J_{ij} - J^*_{ij}| \cdot \left(\frac{1}{2}|J_{ij} - J^*_{ij}| + \frac{1}{2}\sum_{k\neq i,j} |J_{ik} - J^*_{ik}| + |J_{ik} - J^*_{jk}|\right) \\ =& \sum_{i=1}^n \sum_{j\neq i} |J_{ij} - J^*_{ij}| \cdot \left(\frac{1}{4}|J_{ij} - J^*_{ij}| + \frac{1}{2}\sum_{k\neq i,j} |J_{ik} - J^*_{ik}|\right) \\ <& \sum_{i=1}^n \sum_{j\neq i} |J_{ij} - J^*_{ij}| \cdot \left(\frac{1}{2}|J_{ij} - J^*_{ij}| + \frac{1}{2}\sum_{k\neq i,j} |J_{ik} - J^*_{ik}|\right) = \sum_{i=1}^n \frac{1}{2}\left(\sum_{j\neq i} |J_{ij} - J^*_{ij}|\right)^2.\end{aligned}$$

By the Cauchy-Schwarzartz inequality, we know that

$$\left(\sum_{j\neq i} |J_{ij} - J^*_{ij}|\right)^2 \leq n\sum_{j\neq i} |J_{ij} - J^*_{ij}|^2$$

Thus, we can have

$$\sum_{i<j}|J_{ij}-J^*_{ij}|\cdot\left(\frac{1}{2}|J_{ij}-J^*_{ij}|+\frac{1}{2}\sum_{k\neq i,j}|J_{ik}-J^*_{ik}|+|J_{ik}-J^*_{jk}|\right)$$
$$<\sum_{i=1}^n\frac{1}{2}\left(\sum_{j\neq i}|J_{ij}-J^*_{ij}|\right)^2\leq\sum_{i=1}^n\frac{n}{2}\sum_{j\neq i}|J_{ij}-J^*_{ij}|^2=\frac{n}{2}\|J-J^*\|_F^2.$$

So, in summary, we have

$$\begin{aligned}&\mathop{\mathbf{E}}_{X\sim\mu}(X^\top(J-J^*)X)+\mathop{\mathbf{E}}_{X\sim\mu^*}(X^\top(J^*-J)X)\\ \leq&4\sum_{i<j}|J_{ij}-J^*_{ij}|\cdot\left(\|\mu-\mu^*\|_{TV}+\frac{1}{2}|J_{ij}-J^*_{ij}|+\frac{1}{2}\sum_{k\neq i,j}|J_{ik}-J^*_{ik}|+|J_{jk}-J^*_{jk}|\right)\\ \leq&2n\|\mu-\mu^*\|_{TV}\cdot\|J-J^*\|_F+2n\|J-J^*\|_F^2\end{aligned}$$

Because in the beginning we have established

$$4\|\mu-\mu^*\|_{TV}^2=\mathop{\mathbf{E}}_{X\sim\mu}(X^\top(J-J^*)X)+\mathop{\mathbf{E}}_{X\sim\mu^*}(X^\top(J^*-J)X),$$

We have the quadratic formula

$$2\|\mu-\mu^*\|_{TV}^2\leq n\|\mu-\mu^*\|_{TV}\cdot\|J-J^*\|_F+n\|J-J^*\|_F^2,$$

which yields that $\|\mu-\mu^*\|_{TV}\leq n\|J-J^*\|_F$. ■

Appendix F. Discussion of guarantee from Theorem 37

Notice that for Ising model with bounded with, Lemma 6 in Dagan et al. (2021) has established that (also in the proof of Theorem 48) there is a constant (depending only on M) that $\mathbf{E}_{J^*}[\|(J-J^*)X\|]\geq C_M\|J-J^*\|_F^2$ and thus we know that the condition $\mathbf{E}[\|(J-J^*)X\|^2]\leq\varepsilon^2$ implies that $\|J-J^*\|_F^2\leq\varepsilon^2/C_M$. Therefore, the Corollary 7 follows from Theorem 6.

Now, we use an example to show that in some cases, it is strictly stronger. Let $\beta_1,\beta_2>1$ be two numbers and $J^*=\frac{\beta_1}{n}\mathbf{1}\mathbf{1}^\top$, and $J=\frac{\beta_2}{n}\mathbf{1}\mathbf{1}^\top$.

It is well known that (see e.g Ellis (2007)) for Curie Weiss model with inverse temperature $\beta>1$, we have with probability at least $1-\exp(-C\sqrt{n})$ that

$$\frac{1}{n}\sum_{i=1}^n X_i=\pm(2s(\beta)-1)+O(n^{-1/4})\ ,$$

where $s(\beta)$ is the (larger) maximum in the optimization problem

$$s(\beta)=\operatorname*{argmax}_s-s\log(s)-(1-s)\log(1-s)+2\beta s(1-s).$$

Now, the condition given in Theorem 6 for matrices J,J^* says that $\mathbf{E}_{J^*}[\|(J-J^*)X\|^2]\leq\varepsilon^2$. This implies that $\frac{(\beta_1-\beta_2)^2}{n}\mathbf{E}_{J^*}[(\sum_{i=1}^n X_i)^2]\leq\varepsilon^2$, or, $(\beta_1-\beta_2)\lesssim\frac{\varepsilon}{\sqrt{n}\cdot(2s(\beta_1)-1)}$. Therefore, this implies $\|J-J^*\|_F\lesssim\frac{\varepsilon}{\sqrt{n}}$, stricter than $\|J-J^*\|_F\lesssim\varepsilon$.

Appendix G. Verifying the regularity condition

In this section, we will demonstrate how the regularity condition of Theorem 37 can be verified whenever our model satisfies two natural high-temperature conditions. These include the following two cases: (1) Dobrushin's condition holds, (2) the model is spectrally bounded. We also show that it can be true even in low temperature settings, with the notable example being the Curie-Weiss model.

We start with the first claim.

Theorem 48 *Suppose $J^* \in \mathcal{S}_0^n$ satisfies $\|J^*\|_\infty < 1-\delta$, where $\delta > 0$. Then, there is some $\epsilon_0 > 0$ that depends on δ, such that the Ising model with interaction matrix J^* satisfies (ϵ_0^2, C)-regularity, where C depends on δ and ϵ.*

Proof We compare $\mathbf{E}[\|(J-J^*)X\|^2]$ with $\|(J-J^*)X\|_F^2$. Specifically, because $\|J^*\|_\infty < 1-\delta$, we can show that, by Lemma 6 in Dagan et al. (2021), there exists a universal constant C_1 such that for all $a \in \mathbb{R}^n$, $\mathbf{E}_{J^*}[\|a^\top X\|] = \mathbf{Var}[a^\top X] \geq C_1\|a\|^2$. Therefore, we know that $\mathbf{E}[\|(J-J^*)X\|^2] \leq \varepsilon_0^2$ implies that

$$\|J-J^*\|_F^2 \leq C_1^{-1} \mathop{\mathbf{E}}_{J^*}[\|(J-J^*)X\|^2] \leq C_1^{-1}\epsilon_0^2 \ ,$$

which then implies that $\|J-J^*\|_\infty \leq \sqrt{C_1^{-1}}\varepsilon_0$. So, we know that $\|J\|_\infty \leq 1-\delta+\sqrt{C_1^{-1}}\varepsilon_0$. Thus, if we take $\varepsilon_0 \leq \sqrt{C_1}\delta/2$, we know that $\|J\|_\infty \leq 1-\delta/2$.

Then, by Theorem 3.7 in Adamczak et al. (2019), we know that for the Ising model with interaction matrix J, there exists a constant C_2 such that the Poincaré inequality holds with coefficient $\frac{C_2}{\delta}$. Therefore, we know that

$$\mathop{\mathbf{E}}_{J}[\|(J-J^*)X\|^2] = \sum_{i=1}^n \mathbf{Var}[(J-J^*)_i^\top X] \leq \frac{C_2}{\delta}\sum_{i=1}^n \|(J-J^*)\|_2^2 = \frac{C_2}{\delta}\|J-J^*\|_F^2$$

We conclude that

$$\mathop{\mathbf{E}}_{J}[\|(J-J^*)X\|^2] \leq \frac{C_2}{\delta}\|J-J^*\|_F^2 \leq \frac{C_2}{\delta C_1}\cdot \mathop{\mathbf{E}}_{J^*}[\|(J-J^*)X\|^2].$$

■

Theorem 49 *Let $J^* \in \mathcal{S}_0^n$ satisfy $\lambda_{\max}(J^*) - \lambda_{\min}(J^*) < 1-\delta$. Then, there exists $\epsilon_0 > 0$ that depends on δ, such that the Ising model with interaction matrix J^* satisfies (ϵ_0^2, C)-regularity, where $C > 0$ is a constant that depends on δ.*

Proof Like the previous lemma, we compare $\mathbf{E}[\|(J-J^*)X\|^2]$ with $\|(J-J^*)X\|_F^2$. Specifically, because $\lambda_{\max}(J^*) - \lambda_{\min}(J^*) < 1-\delta$, we can show that, by Lemma 22, we have

$$\|J-J^*\|_F^2 \leq 2\cosh\left(4\sqrt{\frac{1}{1-\|J^*\|_{op}}}\right)^2 \varepsilon_0^2 \leq 2\cosh\left(4\sqrt{\frac{1}{\delta}}\right)^2 \varepsilon_0^2,$$

which then implies that $\|J-J^*\|_F \leq \sqrt{2}\cosh(4\sqrt{1/\delta})\varepsilon_0$. So, we know that

$$\lambda_{\max}(J^*) - \lambda_{\min}(J^*) \leq 1-\delta+2\sqrt{2}\cosh(4\sqrt{1/\delta})\varepsilon_0 \ .$$

Then, if we take $\varepsilon_0 \le \delta/(4\sqrt{2}\cosh(4\sqrt{1/\delta}))$, we know that $\lambda_{\max}(J^*) - \lambda_{\min}(J^*) \le 1 - \delta/2$.

We can then apply Lemma 15, which says that there exists an absolute constant C_2, such that the Ising model with interaction matrix J satisfies the Poincare inequality with coefficient $\frac{C_2}{\delta}$. Therefore, we know that

$$\mathop{\mathbf{E}}_{J}[\|(J - J^*)X\|^2] \le \frac{C_2}{\delta}\|J - J^*\|_F^2$$

We conclude that

$$\mathop{\mathbf{E}}_{J}[\|(J - J^*)X\|^2] \le \frac{C_2}{\delta}\|J - J^*\|_F^2 \le \frac{2C_2}{\delta}\cosh\left(4\sqrt{\frac{1}{1 - \|J^*\|_{op}}}\right)^2 \mathop{\mathbf{E}}_{J^*}[\|(J - J^*)X\|^2] \ .$$

This finishes the proof. ■

We conclude the section by showing that the Curie-Weiss model satisfies the regularity condition even at low temperatures.

Theorem 50 *Let J^* be the matrix for the Curie-Weiss model. That is, $J^* = \frac{\beta}{n}(J - I)$ where J is the all-one matrix and $\beta > 1$. Then, we have J^* satisfies $(c_1/n, c_2)$ regularity, where c_1, c_2 are constants that depends on β.*

Proof We know that the distribution is symmetric for all X_i. Therefore, the covariance matrix of $X_1, \ldots, X_n$, denoted by Σ, is a matrix of the form $aJ + (1-a)I$ for some $0 < a < 1$. By Deb and Mukherjee (2023), we can calculate that a converges to $(2s-1)^2$ where $s > 1/2$ is the solution of the equation $2(2s-1) = \log(\frac{s}{1-s})$. Therefore, for any A, we can calculate that

$$\mathop{\mathbf{E}}_{J^*}[\|AX\|^2] = \mathrm{Tr}(A\,\mathbf{E}[XX^\top]A) = \mathrm{Tr}(A(aI + (1-a)J)A) = (1-a)\|A\|_F^2 + a\|A\mathbf{1}\|^2$$

Now, suppose $\mathbf{E}_{J^*}[\|AX\|^2] = \frac{\varepsilon}{n}$, where $\varepsilon < c_1$. We will prove that

$$\mathop{\mathbf{E}}_{J}[\|AX\|^2] \le c_2 \cdot \mathop{\mathbf{E}}_{J^*}[\|AX\|^2] \ ,$$

for some constant $c_2 > 0$.

First of all, by assumption we have $a \cdot (\|A\mathbf{1}\|^2) + (1-a)\|A\|_F^2 = \frac{\varepsilon}{n}$. Therefore, we have $\|A\|_{op} \le \|A\|_F \le \sqrt{\frac{\varepsilon}{(1-a)n}}$, and $\|A\mathbf{1}\| \le \sqrt{\frac{\varepsilon}{an}}$.

Also, we can calculate the following:

$$\mathop{\mathbf{E}}_{J^*}[X^\top AX] = \mathrm{Tr}(A\,\mathbf{E}[XX^\top]) = \mathrm{Tr}(A(aI + (1-a)J)) = (1-a)\mathbf{1}^\top A\mathbf{1} \ ,$$

since A has zero diagonal.

Thus, we can derive that

$$|\mathbf{1}^\top A\mathbf{1}| \le \sqrt{n} \cdot \|A\mathbf{1}\| \le \sqrt{n} \cdot \sqrt{\frac{\varepsilon}{an}} = \sqrt{\varepsilon/a}.$$

The quantity $X^\top AX$ is important, since it represents the difference in Hamiltonians of J and J^*.

First, we show an inequality: for some universal constant c (without loss of generality $c \leq 1$)

$$\Pr_{J^*}\left[\left|X^\top AX - |\mathbf{1}^\top A\mathbf{1}|\right| \geq t\right] \leq 2\exp\left(-c\min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|_{op}}\right)\right)$$

This means that the difference in Hamiltonians is upper-bounded with high probability.

To do that, we can use Deb and Mukherjee (2023) and the Hanson-Wright inequality. By Proposition 4.2 in Deb and Mukherjee (2023), the Curie Weiss model can be written as a mixture of i.i.d. distributions of X_i. Importantly, each of these mixtures has the property that all variables have the same mean. Therefore, by Hanson-Wright inequality for each product measure μ, there exists a universal constant c (for all μ) such that

$$\Pr_{\mu}[|X^\top AX - \mathop{\mathbf{E}}_{\mu}[X^\top AX]| \geq t] \leq 2\exp\left(-c\min\left(\frac{t^2}{\|A\|_F^2 + [\mathbf{E}_\mu \|AX\|]^2}, \frac{t}{\|A\|_{op}}\right)\right).$$

Therefore, if we denote $b_\mu = 1 - \mathbf{E}_\mu[X_i^2]$ (remember this value is the same for all i by Deb and Mukherjee (2023)), then

$$\mathop{\mathbf{E}}_{\mu}[X^\top AX] = \mathrm{Tr}(A\mathop{\mathbf{E}}_{\mu}[XX^\top]) = \mathrm{Tr}(A\mathop{\mathbf{E}}_{\mu}(b_\mu I + (1-b_\mu)J)) \ .$$

Therefore, we have for all i.i.d. measures μ in the decomposition

$$\mathop{\mathbf{E}}_{\mu}[X^\top AX] = \mathrm{Tr}(A(b_\mu I + (1-b_\mu)J)) = (1-b_\mu)\mathrm{Tr}(AJ) = (1-b_\mu)(\mathbf{1}^\top A\mathbf{1}) \leq |\mathbf{1}^\top A\mathbf{1}|,$$

Also, we can calculate that for any i.i.d. μ, $\mathbf{E}_\mu[XX^\top] = b_\mu J + (1-b_\mu)I \preceq I + J$ and therefore,

$$\max_{\mu} \mathop{\mathbf{E}}_{\mu}[\|AX\|^2] \leq \|A\mathbf{1}\|^2 + \|A\|_F^2 \leq \frac{\varepsilon/n}{a(1-a)} \ ,$$

where the maximum is with respect to all product measures in the decomposition. Let $u = |\mathbf{1}^\top A\mathbf{1}|$ and $v = \frac{\varepsilon/n}{a(1-a)}$. Then, for all μ, by Hanson-Wright inequality

$$\Pr_{\mu}[X^\top AX - |\mathbf{1}^\top A\mathbf{1}| \geq t] \leq \Pr_{\mu}[X^\top AX - \mathop{\mathbf{E}}_{\mu}[X^\top AX] \geq t] \leq 2\exp\left(-c\min\left(\frac{t^2}{\|A\|_F^2 + v}, \frac{t}{\|A\|_{op}}\right)\right) \ .$$

Taking expectation with respect to the random choice of measure μ in the decomposition, we can derive the tail bound

$$\Pr_{J^*}[X^\top AX - |\mathbf{1}^\top A\mathbf{1}| \geq t] \leq 2\exp\left(-c\min\left(\frac{t^2}{\|A\|_F^2 + v}, \frac{t}{\|A\|_{op}}\right)\right) \ ,$$

which is what we wished to show. Similarly, we can prove for the lower tail,

$$\Pr_{J^*}[X^\top AX + |\mathbf{1}^\top A\mathbf{1}| \leq -t] \leq 2\exp\left(-c\min\left(\frac{t^2}{\|A\|_F^2 + v}, \frac{t}{\|A\|_{op}}\right)\right).$$

Now let $A = J - J^*$, so we can write that

$$\mathop{\mathbf{E}}_{J}[\|AX\|^2] = \mathop{\mathbf{E}}_{J^*}\left[\frac{Z(J^*)\exp(\frac{1}{2}X^\top AX)}{Z(J)}\|AX\|^2\right].$$

We first lower bound $\frac{Z(J)}{Z(J^*)}$. We have

$$\frac{Z(J)}{Z(J^*)} = \mathop{\mathbf{E}}_{J^*}\left[\exp(\frac{1}{2}X^\top AX)\right].$$

We note that, for a bounded random variable Z and a differentiable increasing function $f(x)$ such that $f(0) = 1$, if we want to lower bound $\mathbf{E}[e^Z]$, we can do the following: let $Y = \min(Z, 0)$. Therefore, we have

$$\begin{aligned}\mathbf{E}[f(Z)] \geq \mathbf{E}[f(Y)] = \mathbf{E}[1 - \int_Y^0 f'(t)\mathrm{d}t] = 1 - \mathbf{E}[\int_{-\infty}^0 \mathbb{1}[t \geq Y]f'(t)\mathrm{d}t] \\ = 1 - \int_{-\infty}^0 f'(t)\,\mathbf{E}[\mathbb{1}[t \geq Y]]\mathrm{d}t = 1 - \int_{-\infty}^0 f'(t)\,\mathbf{Pr}[Z \leq t]\mathrm{d}t.\end{aligned}$$

The second line is by Fubini. By the above inequality and the tail bound, when we view Z to be $X^\top AX + u$ and $f(x) = e^{x/2}$, we can calculate that

$$\begin{aligned}&\mathop{\mathbf{E}}_{J^*}\left[\exp\left(\frac{1}{2}X^\top AX\right)\right] = e^{-u/2}\mathop{\mathbf{E}}_{J^*}\left[\exp(\frac{1}{2}(X^\top AX + u))\right] \\ \geq& e^{-u/2}\cdot\left(1 - \frac{1}{2}\int_0^\infty e^{-t/2}\mathop{\mathbf{Pr}}_{J^*}[X^\top AX + u \leq -t]\mathrm{d}t\right) \\ \geq& e^{-u/2}\cdot\left(1 - \int_0^\infty \exp\left(-\frac{t}{2} - c\min\left(\frac{t^2}{\|A\|_F^2 + v}, \frac{t}{\|A\|_{op}}\right)\right)\mathrm{d}t\right).\end{aligned}$$

Since we know $\|A\|_{op} \leq \|A\|_F \leq v$, we have that

$$\begin{aligned}&\mathop{\mathbf{E}}_{J^*}[\exp(\frac{1}{2}X^\top AX)] = e^{-u/2}\mathop{\mathbf{E}}_{J^*}\left[\exp\left(\frac{1}{2}(X^\top AX + u)\right)\right] \\ \geq& e^{-u/2}\cdot\left(1 - \int_0^\infty \exp\left(-\frac{t}{2} - c\min\left(\frac{t^2}{\|A\|_F^2 + v}, \frac{t}{\|A\|_{op}}\right)\right)\mathrm{d}t\right) \\ \geq& e^{-u/2}\cdot\left(1 - \int_0^\infty \exp\left(-\frac{t}{2} - c\min\left(\frac{t^2}{2v}, \frac{t}{\sqrt{v}}\right)\right)\mathrm{d}t\right) \\ =& e^{-u/2}\cdot\left(1 - \int_0^{2\sqrt{v}} \exp\left(-\frac{t}{2} - c\frac{t^2}{2v}\right)\mathrm{d}t - \int_{2\sqrt{v}}^\infty \exp\left(-\frac{t}{2} - c\frac{t}{\sqrt{v}}\right)\mathrm{d}t\right) \\ \geq& e^{-u/2}\cdot\left(1 - \int_0^{2\sqrt{v}} 1\mathrm{d}t - \int_0^\infty \exp\left(-\frac{t}{2} - c\frac{t}{\sqrt{v}}\right)\mathrm{d}t\right) \\ =& e^{-u/2}\cdot\left(1 - 2\sqrt{v} - \frac{1}{c/\sqrt{v} + 1/2}\right).\end{aligned}$$

We choose the ε small enough such that $\sqrt{v} = \sqrt{\frac{\varepsilon}{na(1-a)}} \leq c/8$, so we have

$$\begin{aligned}&\mathop{\mathbf{E}}_{J^*}[\exp(X^\top AX)] \geq e^{-u/2}\cdot\left(1 - \|A\|_F - \frac{1}{c/\|A\|_F + 1/2}\right) \\ \geq& e^{-u/2}\cdot(1 - 1/4 - 1/4) \geq e^{-\sqrt{\varepsilon/a}/2}/2.\end{aligned}$$

Finally, we upper bound the expectation as follows

$$\mathop{\mathbf{E}}_{J^*}\left[\exp\left(\frac{1}{2}X^\top AX\right)\|AX\|^2\right] \le \mathop{\mathbf{E}}_{J^*}\left[\exp\left(\frac{\sqrt{n}}{2}\|AX\|\right)\|AX\|^2\right]$$

Using the Hanson-Wright inequality as before, replacing A with A^2, we have

$$\mathop{\mathbf{Pr}}_{J^*}[\|AX\|^2 - \max_\mu \mathop{\mathbf{E}}_\mu[\|AX\|^2] \ge t] \le 2\exp\left(-c\min(\frac{t^2}{\|A^2\|_F^2 + \max_\mu \mathbf{E}_\mu[\|A^2X\|]^2}, \frac{t}{\|A^2\|_{op}})\right)$$
$$\le 2\exp\left(-c\min\left(\frac{t^2}{\|A\|_F^4 + \|A\|_F^2 \max_\mu \mathbf{E}_\mu[\|AX\|^2]}, \frac{t}{\|A\|_F^2}\right)\right).$$

Here, as before, $\max_\mu$ is taking $\max$ function for all i.i.d. measure in the decomposition of the Curie Weiss model Deb and Mukherjee (2023).

Similar to the proof of the lower bound from before, if Z is a bounded random variable and f is an increasing differentiable function such that $f(0) = 0$, if we want to upper bound $\mathbf{E}[f(Z)]$, we can let $W = \max(Z, 0)$ and have

$$\mathbf{E}[f(Z)] \le \mathbf{E}[f(W)] = \mathbf{E}[\int_0^W f'(t)\mathrm{d}t] = \mathbf{E}[\int_0^\infty \mathbb{1}[t \le W]f'(t)\mathrm{d}t]$$
$$= \int_0^\infty f'(t)\,\mathbf{E}[\mathbb{1}[W \ge t]]\mathrm{d}t = \int_0^\infty f'(t)\,\mathbf{Pr}[Z \ge t]\mathrm{d}t.$$

Let $f(t) = te^{\sqrt{nt}/2}$, and $f'(t) = \exp(\frac{\sqrt{nt}}{2})(1 + \frac{\sqrt{nt}}{2})$. The random variable Z will be $\|AX\|^2$. We will now use the concentration result for $\|AX\|^2$ to upper bound the expectation. Below, we will use the simple observation that we made before $\max_\mu \mathbf{E}_\mu[\|AX\|^2] \le v$. Using the identity $\exp(\frac{\sqrt{nt}}{2})(1 + \frac{\sqrt{nt}}{2}) \le \exp(\sqrt{nt})$ in the rest of the calculation, we can calculate that

$$\mathop{\mathbf{E}}_{J^*}\left[\exp\left(\frac{\sqrt{n}}{2}\|AX\|\right)\|AX\|^2\right]$$
$$\le \int_0^\infty \exp\left(\frac{\sqrt{nt}}{2}\right)\left(1 + \frac{\sqrt{nt}}{2}\right)\mathbf{Pr}(\|AX\|^2 \ge t)\mathrm{d}t$$
$$\le \int_0^\infty \exp\left(\sqrt{nt}\right)\mathbf{Pr}(\|AX\|^2 \ge t)\mathrm{d}t$$
$$\le \int_0^{2v+\|A\|_F^2} \exp(\sqrt{nt})\mathrm{d}t + \int_{2v+\|A\|_F^2}^\infty \exp(\sqrt{nt}) \cdot 2\exp\left(-c\min\left(\frac{(t-v)^2}{\|A\|_F^4 + \|A\|_F^2 v}, \frac{t-v}{\|A\|_F^2}\right)\right)\mathrm{d}t$$
$$= \int_0^{2v+\|A\|_F^2} \exp(\sqrt{nt})\mathrm{d}t + \int_{2v+\|A\|_F^2}^\infty \exp(\sqrt{nt}) \cdot 2\exp\left(-c\frac{t-v}{\|A\|_F^2}\right)\mathrm{d}t$$

Here, the first inequality is by taking Z to be $\|AX\|^2$ and $f(x) = xe^{\sqrt{nx}}$. The second one is because of the inequality we mentioned, the third one is by the following: if $t \le 2v + \|A\|_F^2$, the probability of $\|AX\|^2 \ge t$ is at most 1. Otherwise, we use the concentration tail bound and get the upper bound.

We know that $nv = \varepsilon/(a(1-a))$, and $\|A\|_F^2 \le \varepsilon/(n(1-a)) \le v$, so we have

$$\int_0^{2v+\|A\|_F^2} \exp(\sqrt{nt})\mathrm{d}t \le (2v + \|A\|_F^2)e^{\sqrt{n(2v+\|A\|_F^2)}} \le \frac{3\varepsilon}{a(1-a)n}\exp\left(\frac{3\varepsilon}{a(1-a)}\right).$$

Also, if we take $\varepsilon < 1 - a$, we have

$$\int_{2v+\|A\|_F^2}^{\infty} \exp\left(\sqrt{nt}\right) \cdot 2\exp\left(-c\frac{t-v}{\|A\|_F^2}\right) \mathrm{d}t \leq \int_{v+\|A\|_F^2}^{\infty} \exp(\sqrt{nt}) \cdot 2\exp\left(-c\frac{t-v}{\|A\|_F^2}\right) \mathrm{d}t$$
$$\leq \int_{v+\|A\|_F^2}^{\infty} \exp(1/(4c) + cnt) \cdot 2\exp\left(-c\frac{t-v}{\|A\|_F^2}\right) \mathrm{d}t$$
$$= \exp\left(\frac{1}{4c} + (v + \|A\|_F^2) \cdot c \cdot n - \|A\|_F^2 \cdot \frac{c}{\|A\|_F^2}\right) \cdot \frac{1/c}{\frac{1}{\|A\|_F^2} - n}$$
$$\leq \exp\left(\frac{1}{4c} + (v + \|A\|_F^2)cn - c\right) \cdot \frac{1/c}{1-a-\varepsilon} \cdot \frac{\varepsilon}{n} \leq \exp\left(\frac{1}{4c} + \frac{\varepsilon \cdot (2-a)}{a(1-a)} - c\right) \cdot \frac{1/c}{1-a-\varepsilon} \cdot \frac{\varepsilon}{n}.$$

Therefore, summing up all the things above, we have

$$\mathop{\mathbf{E}}_{J}[\|AX\|^2] \leq \frac{\varepsilon}{n}\left(\frac{3}{a(1-a)}\exp\left(\frac{3\varepsilon}{a(1-a)}\right) + \exp\left(\frac{1}{4c} + \frac{\varepsilon \cdot (2-a)}{a(1-a)} - c\right) \cdot \frac{1/c}{1-a-\varepsilon}\right) \cdot 2e^{\sqrt{\varepsilon/a}/2}.$$

Since a is a constant depending on β, and we can take ε as a sufficiently small constant depending on β, the result follows. ■