The Annals of Statistics
2020, Vol. 48, No. 2, 785–810
https://doi.org/10.1214/19-AOS1822
© Institute of Mathematical Statistics, 2020

JOINT ESTIMATION OF PARAMETERS IN ISING MODEL

BY PROMIT GHOSAL* AND SUMIT MUKHERJEE**

*Department of Statistics, Columbia University, *pg2475@columbia.edu; **sm3949@columbia.edu*

We study joint estimation of the inverse temperature and magnetization parameters (β, B) of an Ising model with a nonnegative coupling matrix A_n of size $n \times n$, given one sample from the Ising model. We give a general bound on the rate of consistency of the bi-variate pseudo-likelihood estimator. Using this, we show that estimation at rate $n^{-1/2}$ is always possible if A_n is the adjacency matrix of a bounded degree graph. If A_n is the scaled adjacency matrix of a graph whose average degree goes to $+\infty$, the situation is a bit more delicate. In this case, estimation at rate $n^{-1/2}$ is still possible if the graph is not regular (in an asymptotic sense). Finally, we show that consistent estimation of both parameters is impossible if the graph is Erdős–Renyi with parameter $p > 0$ independent of n, thus confirming that estimation is harder on approximately regular graphs with large degree.

1. Introduction. Suppose $\beta > 0$, $B \neq 0$ are unknown parameters, and A_n is an $n \times n$ symmetric matrix with nonnegative entries with 0 on the diagonal. For $\mathbf{x} := (x_1, \dots, x_n) \in \{-1, 1\}^n$, define a p.m.f. $\mathbb{P}_{n,\beta,B}(\cdot)$ by setting

$$\mathbb{P}_{n,\beta,B}(\mathbf{X} = \mathbf{x}) = \frac{1}{Z_n(\beta, B)} e^{\frac{\beta}{2}\mathbf{x}' A_n \mathbf{x} + B \sum_{i=1}^n x_i}. \tag{1.1}$$

This is the Ising model with coupling matrix A_n, and inverse temperature parameter β and magnetization parameter B. Study of Ising models is a growing area which has received significant attention in statistics and machine learning in recent years. In this paper, we focus on estimation in Ising models, the existing literature on which can be broadly classified into two categories. One of the branches assumes that the matrix A_n is the unknown parameter of interest, and focuses on estimating A_n under the assumption that i.i.d. copies $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(p)}$ are available from the model described in (1.1) (cf. [1, 7, 20, 21] and references therein). Another branch works under the assumption that only one observation $\mathbf{X}$ is available from the model in (1.1) (cf. [5, 8, 10, 11, 16, 17] and references therein). In this setting, estimation of the whole matrix A_n (which has n^2 entries) is impossible from a vector $\mathbf{X}$ of size n (without making "strong" assumptions on A_n). As such, the standard assumption is that the matrix A_n is completely specified, and the focus is on estimating the parameters (β, B). In this direction, the behavior of the MLE for the Curie–Weiss model (when A_n is the scaled adjacency matrix of the complete graph) was studied in [11], where the authors showed that in the regime $\beta > 0$, $B \neq 0$, the MLE of β is $\sqrt{n}$ consistent for β if B is known, and vice versa. They also show that if both (β, B) are unknown, then the joint MLE for the model does not exist with probability 1. This raises the natural question as to whether there are other estimators which work in this case. Focusing on the case when $B = 0$ is known, [8] gave general sufficient conditions under which the pseudo-likelihood estimator (defined below) is a $\sqrt{n}$ consistent estimator. The pseudo-likelihood estimator is a natural estimator for such models, as computing the pseudo-likelihood estimator does not require knowledge of the partition function $Z_n(\beta, B)$, as opposed to computing the Maximum Likelihood Estimate (MLE). Extending

Received February 2018; revised October 2018.
MSC2010 subject classifications. Primary 62F12; secondary 60F10.
Key words and phrases. Ising model, pseudo-likelihood.

the ideas of [5, 8] studies the behavior of the rate of consistency of the pseudo-likelihood estimator at all values of β, demonstrating interesting phase transition properties in the rate of the pseudo-likelihood estimator. The question of joint estimation of (β, B) for a general matrix A_n was raised in [8]. To the best of our knowledge, this question has not been addressed in the literature. The introduction and study of the second parameter B is natural, as B controls the expected value of $\sum_{i=1}^n X_i$, and allows for more flexible modeling of data. The problem of inference of both parameters simultaneously turns out to be significantly more challenging. The strong dependence of the variables in Ising models ensure that there is no independence to exploit. Even getting estimates of the means and the variances of the random variables is not an easy task. As such, understanding the criterion obtained for consistency and relating it to easily verifiable properties of the graph sequence require a sequence of technical arguments. In this process, we discover an interesting dichotomy for hardness of the estimation problem, depending on whether the graph is “regular” and/or “dense” (cf. Section 1.2). Such a dichotomy was not present in the one parameter problem.

1.1. *Main results.* Throughout this paper, we will assume that (β, B) are unknown parameters of interest. Let us begin by introducing the bivariate pseudo-likelihood estimator for (β, B).

DEFINITION. For any $i \in [n]$, we have

$$\mathbb{P}_{n,\beta,B}(X_i = 1 | X_j, j \neq i) = \frac{e^{\beta m_i(\mathbf{x})+B}}{e^{\beta m_i(\mathbf{x})+B} + e^{-\beta m_i(\mathbf{x})-B}},$$

where $m_i(\mathbf{x}) := \sum_{j=1}^n A_n(i,j)x_j$. Define the pseudo-likelihood as the product of the one-dimensional conditional distributions (see [3, 4]):

$$\prod_{i=1}^n \mathbb{P}_{n,\beta,B}(X_i = x_i | X_j, j \neq i)$$
$$= 2^{-n} \exp\left\{\sum_{i=1}^n \bigl(\beta x_i m_i(\mathbf{x}) + B x_i - \log\cosh\bigl(\beta m_i(\mathbf{x}) + B\bigr)\bigr)\right\}.$$

On taking log and differentiating this with respect to (β, B), we get the vector $(Q_n(\beta, B|\mathbf{x}), R_n(\beta, B|\mathbf{x}))$, where

$$Q_n(\beta, B|\mathbf{x}) := \sum_{i=1}^n m_i(\mathbf{x})\bigl(x_i - \tanh\bigl(\beta m_i(\mathbf{x}) + B\bigr)\bigr),$$
$$R_n(\beta, B|\mathbf{x}) := \sum_{i=1}^n \bigl(x_i - \tanh\bigl(\beta m_i(\mathbf{x}) + B\bigr)\bigr).$$

The bivariate equation

$$\mathrm{PL}_n(\beta, B|\mathbf{x}) := \bigl(Q_n(\beta, B|\mathbf{x}), R_n(\beta, B|\mathbf{x})\bigr) = (0, 0)$$

will be referred to as the pseudo-likelihood equation in this paper. If the pseudo-likelihood equation has a unique root in $(\beta, B) \in \mathbb{R}^2$, denote it by $(\hat{\beta}_n, \hat{B}_n)$. This is the pseudo-likelihood estimator for the parameter vector (β, B).

We will need some assumptions on the coupling matrix A_n for the analysis of the pseudo-likelihood estimator. Throughout this paper, we assume that A_n has nonnegative entries and is completely known. For any $i \in [n]$, let $\mathcal{R}_n(i) := \sum_{j=1}^n A_n(i,j)$ denote the ith row sum of A_n, and let $\bar{\mathcal{R}}_n := \frac{1}{n}\sum_{i=1}^n \mathcal{R}_n(i)$ denote the average of the row sums. Assume further that

$$\max_{i\in[n]} \mathcal{R}_n(i) \le \gamma, \tag{1.2}$$

$$\liminf_{n\to\infty} \bar{\mathcal{R}}_n > 0. \tag{1.3}$$

Here, γ is a finite constant independent of n. Note that (1.2) implies that A_n satisfies the following condition:

$$\sup_{\mathbf{x}\in[0,1]^n} \sum_{i=1}^n \left| \sum_{j=1}^n A(i,j)x_j \right| = O(n) \tag{1.4}$$

of [2], equation (1.10), as well as $\|A_n\|_2 \le \gamma$ ([8], Condition (a), Theorem 1.1), where $\|A_n\|_2$ is the operator norm of A_n. If (1.4) does not hold, the log-partition function $\log Z_n(\beta, B)$ grows super linearly, thus giving $\lim_{n\to\infty} \frac{1}{n}\log Z_n(\beta,B) = +\infty$ via mean field lower bound. On the other hand, the condition $\|A_n\|_2 \le \gamma$ is a regularity condition which ensures that no eigenvalue has an unduly large effect on the corresponding Ising model. Both (1.4) and $\|A_n\|_2 \le \gamma$ are satisfied by "almost all" Ising models studied in the literature. Condition (1.3) is of a different spirit, which ensures that the resulting Ising model is nontrivial to rule out cases when A_n is close to the $\mathbf{0}$ matrix.

Stating our results requires the following definition.

DEFINITION. Suppose U_n and V_n are two nonnegative random variables on the probability space $(\{-1,1\}^n, \mathbb{P}_{n,\beta,B})$, where $\mathbb{P}_{n,\beta,B}$ is the Ising p.m.f. given in (1.1). We will say $U_n = O_p(V_n)$ if the sequence $\frac{U_n}{V_n}$ is tight. In particular, this implies that

$$\lim_{n\to\infty} \mathbb{P}_{n,\beta,B}(U_n > 0, V_n = 0) = 0.$$

We will say $U_n = \Theta_p(V_n)$, if both $U_n = O_p(V_n)$ and $V_n = O_p(U_n)$. We will say $U_n = o_p(V_n)$ if $\frac{U_n}{V_n} \xrightarrow{p} 0$. We remove the subscript p from Θ_p, O_p and o_p if U_n and V_n are deterministic sequences of positive reals.

DEFINITION 1.1. Set $\bar{m}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n m_i(\mathbf{x})$ and $T_n(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n (m_i(\mathbf{x}) - \bar{m}(\mathbf{x}))^2$.

Our first result gives a general upper bound on the error of the pseudo-likelihood estimator.

THEOREM 1.2. *Suppose* $\mathbf{X} = (X_1, \dots, X_n)$ *is an observation from the Ising model* (1.1), *where the coupling matrix* A_n *satisfies* (1.2) *and* (1.3), *and* $\beta > 0$, $B \ne 0$.

(a) *For every* $\mathbf{x} \in \{-1,1\}^n$, *the pseudo-likelihood estimator* $(\hat{\beta}_n, \hat{B}_n)$ *exists iff* $\mathbf{x} \in A_{1,n}^c \cap A_{2,n}^c \cap A_{3,n}^c \cap A_{4,n}^c$, *where*

$$\begin{aligned}
A_{1,n} &:= \{\mathbf{x} \in \{-1,1\}^n : T_n(\mathbf{x}) = 0\},\\
A_{2,n} &:= \{\mathbf{x} \in \{-1,1\}^n : m_i(\mathbf{x})x_i = |m_i(\mathbf{x})| \textit{ for all } i \in [n]\},\\
A_{3,n} &:= \{\mathbf{x} \in \{-1,1\}^n : m_i(\mathbf{x})x_i = -|m_i(\mathbf{x})| \textit{ for all } i \in [n]\},\\
A_{4,n} &:= \{\mathbf{1}, -\mathbf{1}\}.
\end{aligned}$$

(b) *If the true parameter is* (β_0, B_0), *then we have*

$$\lim_{n\to\infty} \mathbb{P}_{n,\beta_0,B_0}(\mathbf{X} \in A_{2,n}^c \cap A_{3,n}^c \cap A_{4,n}^c) = 1.$$

(c) *Further, if* $\frac{1}{T_n(\mathbf{X})} = o_p(\sqrt{n})$, *then*

$$\|\hat{\beta}_n - \beta_0, \hat{B}_n - B_0\|_2 = O_p\left(\frac{1}{\sqrt{n}T_n(\mathbf{X})}\right).$$

In particular, $(\hat{\beta}_n, \hat{B}_n)$ *is jointly consistent for* (β, B).

An immediate corollary of Theorem 1.2 is the following corollary.

COROLLARY 1.3. *In the setting of Theorem* 1.2, *if we further have*

$$T_n(\mathbf{X}) = \Theta_p(1), \tag{1.5}$$

then $\|\hat{\beta}_n - \beta_0, \hat{B}_n - B_0\|_2 = O_p(\frac{1}{\sqrt{n}})$ *under* $\mathbb{P}_{n,\beta_0,B_0}$, *that is, the joint pseudo-likelihood estimator is jointly* $\sqrt{n}$ *consistent.*

Corollary 1.3 shows that (1.5) is a sufficient condition for $\sqrt{n}$ consistency of the pseudo-likelihood estimate. However, the condition (1.5) is an implicit condition, and it is not straightforward to verify (1.5) directly in most examples. To understand (1.5), we separate our analysis into two cases, depending on whether the matrix A_n is "mean field" or not.

DEFINITION. We say that a sequence of matrices $\{A_n\}_{n\geq 1}$ satisfies the mean field condition if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i,j=1}^{n} A_n(i,j)^2 = 0. \tag{1.6}$$

Condition (1.6) was first introduced in [2] to study the limiting behavior of normalizing constant of Ising and Potts models. For examples of matrices which are mean field, we refer the reader to Section 1.2. Our first result of this section now gives a simple sufficient condition for joint $\sqrt{n}$ consistency of the pseudo-likelihood estimator for mean field matrices. Note that (1.7) and (1.2) together imply (1.3), and so it need not be assumed separately.

THEOREM 1.4. *Suppose* $\mathbf{X} = (X_1, \dots, X_n)$ *is an observation from the Ising model* (1.1), *where the coupling matrix* A_n *satisfies* (1.2) *and* (1.6), *and* $\beta > 0$, $B \neq 0$. *If*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\mathcal{R}_n(i) - \bar{\mathcal{R}}_n)^2 > 0, \tag{1.7}$$

then we have $T_n(\mathbf{X}) = \Theta_p(1)$ *under* $\mathbb{P}_{n,\beta,B}$. *Consequently, the pseudo-likelihood estimator is jointly* $\sqrt{n}$ *consistent.*

This raises the natural question as to what happens for mean field matrices when (1.7) does not hold, that is, A_n is asymptotically regular (cf. (1.8)). The following theorem addresses this question by showing that whenever the coupling matrix A_n is mean field and asymptotically regular, the random variable $T_n(\mathbf{X})$ is $o_p(1)$.

THEOREM 1.5. *Suppose* $\mathbf{X} = (X_1, \dots, X_n)$ *is an observation from the Ising model* (1.1), *where the coupling matrix* A_n *satisfies* (1.2) *and* (1.6), *and* $\beta > 0$, $B \neq 0$, *If*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\mathcal{R}_n(i) - \bar{\mathcal{R}}_n)^2 = 0, \tag{1.8}$$

then we have $T_n(\mathbf{X}) = o_p(1)$ *under* $\mathbb{P}_{n,\beta,B}$.

The notion of asymptotic regularity captures those matrices for which the empirical distribution of the row sums $\mathcal{R}_n(i)$ converges in probability to 1. For examples of matrices which are asymptotically regular, we refer the reader to Section 1.2. Theorem 1.5 along with the upper bound of Theorem 1.2 together suggest that $\sqrt{n}$ consistency may not be attained by the pseudo-likelihood estimator for asymptotically regular matrices which satisfies the mean field condition (1.6). In particular, if A_n is the scaled adjacency matrix of an Erdős–Renyi graph with parameter p_n, then the mean field condition (1.6) holds if $np_n \to \infty$, and asymptotic regularity (1.8) holds if $p_n \geq (1+\delta)\frac{\log n}{n}$ with $\delta > 0$ independent of n (cf. Section 1.2). Under the assumption that the parameter $p_n = p \in (0,1]$ of the Erdős–Renyi graph is independent of n, the following theorem shows the much stronger result that there does not exist *any* consistent sequence of estimators for (β, B). Of course, this rules out the possibility of any $\sqrt{n}$-consistent estimators as well in this setting.

THEOREM 1.6. *Suppose* $\mathbf{X} = (X_1, \dots, X_n)$ *is an observation from the Ising model* (1.1), *where the coupling matrix is* $A_n(i,j) = \frac{1}{(n-1)p} G_n(i,j)$, *where* G_n *is a random graph from* $\mathcal{G}(n,p)$, *the Erdős–Renyi graph with parameter* $p > 0$, *independent of* n. *Let* $t \in (0,1)$ *be fixed*, *and let*

$$\Gamma_t := \{(\beta, B) \in (0,\infty)^2 : t = \tanh(\beta t + B)\}.$$

Let $\mathbb{P}^{\text{er}}_{n,\beta,B}$ *denote the joint law of* $(\mathbf{X}, G_n)$ *on* $\{-1,1\}^n \times \{0,1\}^{\binom{n}{2}}$. *Then, setting* $\mathbb{Q}_n$ *to be product measure on* $\{-1,1\}^n$ *under which*

$$\mathbb{Q}_n(X_i = 1) = \frac{1}{1 + e^{-2\tanh^{-1}(t)}},$$

we have that $\mathbb{Q}_n \times \mathcal{G}(n,p)$ *is contiguous to* $\mathbb{P}^{\text{er}}_{n,\beta,B}$ *for every* $(\beta, B) \in \Gamma_t$. *Consequently*, *under* $\mathbb{P}^{\text{er}}_{n,\beta,\mathbb{B}}$ *there does not exist any sequence of estimates* (*functions of* $(\mathbf{X}, G_n)$) *which is consistent for* (β, B) *in* Γ_t.

REMARK 1.7. It was pointed out in [11] that the MLE for (β, B) does not exist for the Curie– Weiss model. The above theorem extends this by showing that consistent estimates do not exist when the underlying graph of the Ising model is Erdős–Renyi. Note that if we set $p = 1$ in the Erdős–Renyi model we get a complete graph on n vertices, which corresponds to the Curie–Weiss model. More generally, we conjecture that there are no $\sqrt{n}$ consistent estimates for both parameters, for a sequence of regular graphs with degree going to $+\infty$.

If (1.6) does not hold, joint estimation of both parameters at rate $\sqrt{n}$ is always possible irrespective of whether the matrix A_n is regular or not, as shown in the following theorem. Note that (1.2) and (1.9) together imply (1.3), so it is not assumed separately.

THEOREM 1.8. *Suppose* $\mathbf{X} = (X_1, \dots, X_n)$ *is an observation from the Ising model* (1.1), *where the coupling matrix* A_n *satisfies* (1.2) *and*

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i,j=1}^{n} A_n^2(i,j) > 0. \tag{1.9}$$

Then for any $\beta > 0$, $B \neq 0$ *we have* $T_n(\mathbf{X}) = \Theta_p(1)$ *under* $\mathbb{P}_{n,\beta,B}$, *and consequently the pseudo-likelihood estimator is jointly* $\sqrt{n}$ *consistent.*

Figure 1 gives a gist of our results on a summary tree.

To complete the picture, we show that if one of the two parameters are known, then the pseudo-likelihood estimator for the other parameter is $\sqrt{n}$ consistent, for all $\beta > 0$, $B \neq 0$. Thus joint estimation is indeed a much harder problem than estimation of the individual parameters. The proof of this proposition appears in Supplement A [15].

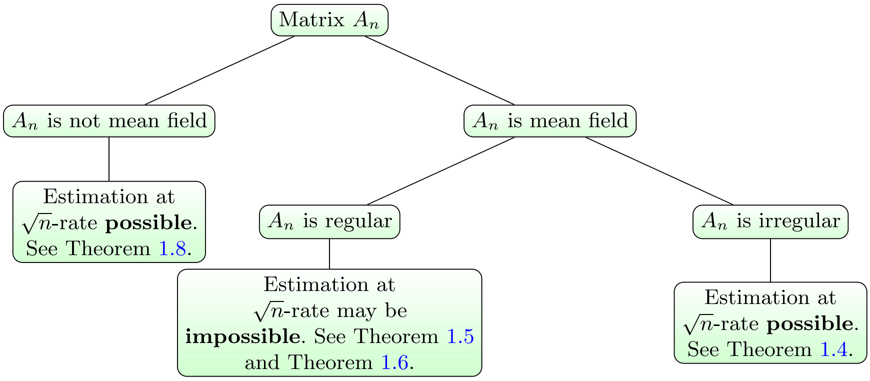
Matrix A_n
A_n is not mean field
A_n is mean field
Estimation at $\sqrt{n}$-rate **possible**. See Theorem 1.8.
A_n is regular
A_n is irregular
Estimation at $\sqrt{n}$-rate may be **impossible**. See Theorem 1.5 and Theorem 1.6.
Estimation at $\sqrt{n}$-rate **possible**. See Theorem 1.4.

FIG. 1. *Summary tree of our results.*

PROPOSITION 1.9. *Suppose* $\mathbf{X} = (X_1, \dots, X_n)$ *is an observation from the Ising model* (1.1), *where the coupling matrix* A_n *satisfies* (1.2) *and* (1.3), *and* $\beta_0 > 0$, $B_0 \neq 0$.

(a) *If* B_0 *is known, then the equation* $Q_n(\beta, B|\mathbf{X}) = 0$ *has a unique root* $\hat{\beta}_n$ *with probability tending to* 1, *which satisfies* $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$ *under* $\mathbb{P}_{n,\beta_0,B}$.

(b) *If* β_0 *is known, then the equation* $R_n(\beta, B|\mathbf{X}) = 0$ *has a unique root* $\hat{B}_n$ *with probability tending to* 1 *which satisfies* $\sqrt{n}(\hat{B}_n - B_0) = O_p(1)$ *under* $\mathbb{P}_{n,\beta,B_0}$.

We would like to point out here that all our arguments are robust enough go through if the true configuration $(\beta_{0,n}, B_{0,n})$ is assumed to depend on n, and converge at any rate to (β_0, B_0) with $\beta_0 > 0$, $B_0 \neq 0$. We avoid doing this case for the sake of notational simplicity, and because no new ideas are needed in this process.

1.2. *Interpretation of results for graphs.* Even though all our results apply for general matrices with nonnegative entries, the most interesting examples for our theorems are the cases when A_n is the scaled adjacency matrix of a simple graph G_n, as defined below.

DEFINITION 1.10. For a graph G_n with vertices labeled by $[n] := \{1, 2, \dots, n\}$, define the coupling matrix A_n by setting

$$A_n(i, j) := \frac{n}{2e(G_n)} 1\{\text{vertices } i \text{ and } j \text{ are connected in } G_n\},$$

where $e(G_n)$ is the number of edges in the graph G_n. Note that this scaling gives $\sum_{i,j=1}^n A_n(i, j)^2 = \frac{n^2}{2e(G_n)}$. Also let $(d_1(G_n), \dots, d_n(G_n))$ denote the labeled degrees of G_n, and let $\bar{d}(G_n)$ denote the average degree of G_n.

This scaling of the adjacency matrix ensures that the resulting Ising model has nontrivial phase transition properties (see, e.g., [2]), which is of much interest in statistical physics and applied probability. The influence of phase transition on inference has received recent attention (cf. [5, 19]). Under this scaling, (1.3) holds trivially, as $\sum_{i,j=1}^n A_n(i, j) = n$, and condition (1.2) demands that $\max_{i\in[n]} d_i(G_n) = O(\bar{d}(G_n))$. Below we give some common examples of graph sequences:

(a) G_n is a bounded degree graph with $\max_{i\in[n]} d_i(G_n) \le M$, and $e(G_n) \ge n\delta$, where M, δ are independent of n. In this case, $\frac{n^2}{2e(G_n)} \ge \frac{n}{M} = \Theta(n)$, which verifies (1.9). Also $\max_{i\in[n]} d_i(G_n) \le M$ and $\bar{d}(G_n) \ge 2\delta$, and so (1.2) holds. Thus Theorem 1.8 concludes that both the parameters (β, B) can be estimated at rate $\sqrt{n}$. In particular, this holds if G_n is a d regular graph with d independent of n. One important example is a sequence of growing k dimensional cubes in Z^k for some $k \in \mathbb{N}$, whence $d = 2k$.

(b) Let G_n be a sequence of d_n regular graphs with $d_n \to \infty$. In this case, $\frac{n^2}{2e(G_n)} = \frac{n}{d_n} = o(n)$, and $\mathcal{R}_n(i) = 1$ for all $i \in [n]$. This verifies (1.2), (1.6) and (1.8). Theorems 1.2 and 1.5 together then suggest that $\sqrt{n}$ consistency may be impossible. Theorem 1.6 confirms this in the special case when $d_n = n - 1$, which corresponds to the complete graph.

(c) Let G_n be an Erdős–Renyi graph with parameter $p_n = p \in (0, 1]$ independent of n. Theorem 1.6 concludes that there are no consistent estimators for (β, B), let alone $\sqrt{n}$ consistency.

(d) More generally, let G_n be an Erdős–Renyi graph with parameter $p_n \geq (1 + \delta)\frac{\log n}{n}$, where $\delta > 0$ is fixed. In this case, $\frac{n^2}{2e(G_n)} = \Theta_p(\frac{1}{p_n}) = o(n)$, and

$$\sum_{i=1}^{n}(\mathcal{R}_n(i) - \bar{\mathcal{R}}_n)^2 \leq \frac{1}{\bar{d}(G_n)^2}\sum_{i=1}^{n}(d_i(G_n) - (n-1)p_n)^2 = \Theta_p\left(\frac{1}{p_n}\right) = o(n).$$

Thus we have verified (1.6) and (1.8). Since $\max_{i\in[n]} d_i(G_n)$ and $\bar{d}(G_n)$ are both $\Theta_p(np_n)$ for p_n in this range, (1.2) holds, and so Theorems 1.2 and 1.5 together suggest that $\sqrt{n}$ consistency may be impossible. Theorem 1.6 confirms this in the special case when $p_n = p \in (0, 1]$ is independent of n.

(e) G_n is a convergent sequence of dense graph converging to the graphon W which is not identically 0 (see [18] for a survey on the literature on graphons/graph limits). If the function $\mathcal{R}(x) := \int_0^1 W(x, y)\,dy$ is not constant almost surely Lebesgue measure, we have

$$\lim_{n\to\infty}\frac{1}{n}(\mathcal{R}_n(i) - \bar{\mathcal{R}}_n)^2 = \frac{\int_0^1(\mathcal{R}(x) - \int_0^1 \mathcal{R}(y)\,dy)^2\,dx}{\int_{[0,1]^2}\mathcal{R}(x, y)\,dx\,dy} > 0,$$

and so (1.7) holds. Also (1.2) and (1.6) are easy to verify for dense graphs, and so Theorem 1.4 shows that estimation of both parameters at rate $\sqrt{n}$ is possible. On the other hand, if $\mathcal{R}(x)$ is a constant almost surely, (1.8) holds. Theorems 1.2 and 1.5 together suggest that $\sqrt{n}$ consistency may be impossible. Theorem 1.6 confirms this in the special case when G_n is an Erdős–Renyi graph, for which $W(\cdot, \cdot) \equiv p$ and $\mathcal{R}(\cdot) \equiv p$.

(f) G_n is a biregular, bipartite graph with bipartition sets $G_{n,1}$ and $G_{n,2}$, with sizes a_n and b_n respectively, and each vertex in G_{n1} has degree c_n, and each vertex in G_{n2} has degree d_n. Also assume that

$$\lim_{n\to\infty}\frac{a_n}{n} = p \in (0, 1).$$

In this case, (1.2) is easy to verify. If $\limsup_{n\to\infty}(c_n + d_n) < \infty$, (1.9) holds, and so Theorem 1.8 concludes that estimation at rate $\sqrt{n}$ is possible. If $\lim_{n\to\infty}(c_n + d_n) = \infty$, then (1.6) holds, and so there are two cases depending on the value of p. If $p \neq \frac{1}{2}$, we have $\mathcal{R}_n(i) = \frac{n}{a_n} \sim \frac{1}{2p}$ for $i \in G_{n1}$, and $\mathcal{R}_n(i) \sim \frac{1}{2(1-p)}$ for $i \in G_{n2}$. This gives

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{R}_n(i) - \bar{\mathcal{R}}_n)^2 \sim \frac{a_n}{n}\left(\frac{1}{2p} - 1\right)^2 + \frac{b_n}{n}\left(\frac{1}{2(1-p)} - 1\right)^2$$

$$\to \frac{(2p-1)^2}{4p(1-p)} > 0.$$

Thus (1.7) holds, and so the pseudo-likelihood estimator is $\sqrt{n}$ consistent by Theorem 1.4. On the other hand, if if $p = \frac{1}{2}$ the graph G_n is asymptotically regular, and so Theorems 1.2 and 1.5 together then suggest that $\sqrt{n}$ consistency may be impossible.

1.3. *Heuristic and proof overview.* To understand the heuristic idea behind Theorem 1.2, recall that the estimating equation for the pseudo-likelihood estimator is $(Q_n(\beta, B), R_n(\beta, B)) = (0,0)$. Both the quantities on the RHS above are shown to be $O_p(\sqrt{n})$ by existing estimates in the literature. Thus a sufficient criterion for $\sqrt{n}$ consistency is that both the eigenvalues of the Hessian matrix are at least Cn with high probability for some $C > 0$. A direct calculation using the determinant of the Hessian shows that the eigenvalues of the Hessian are at least Cn iff $T_n(\mathbf{X}) = \Theta_p(1)$, where $T_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n (m_i(\mathbf{x}) - \bar{m}(\mathbf{x}))^2$ as before. It thus suffices to focus our attention on understanding the behavior of $T_n(\mathbf{X})$, for which we give the following simple condition:

$$T_n(\mathbf{X}) = o_p(1) \text{ iff the graph is approximately regular,}$$
$$\text{with average degree going to } \infty.$$

This implies that for any other class of graphs the pseudo-likelihood estimator is $\sqrt{n}$ consistent. To understand why $T_n(\mathbf{X})$ is expected to be small for dense regular graphs, consider the following heuristic:

If the coupling matrix A_n is the adjacency matrix (modulo some scaling) of a regular graph with degree d large, by the weak law of large numbers, one can expect

$$m_i(\mathbf{X}) = \frac{1}{d}\sum_{j\sim i} X_j \approx \mathbb{E}_{n,\beta,B}\frac{1}{d}\sum_{j\sim i} X_j = \mathbb{E}_{n,\beta,B} m_i(\mathbf{X})$$

with high probability. One the other hand, owing to regularity of the graph it is also expected that $\mathbb{E}_{n,\beta,B} m_i(\mathbf{X})$ is the same for all i. Thus for regular graphs with large degree, one expects that the vector $\{m_i(\mathbf{X}), 1 \le i \le n\}$ is nearly constant with high probability, or equivalently the random variable $T_n(\mathbf{X})$ is small.

To make the above arguments rigorous, we take recourse to the mean field approach developed in [2] to prove a concentration result for the vector $(m_1(\mathbf{X}), \dots, m_n(\mathbf{X}))$ for graphs with large average degree in the Appendix (see Supplement A). Essentially, the concentration result shows that the vector $\mathbf{m}(\mathbf{x})$ (or equivalently, $b(\mathbf{x}) = \tanh(\beta m(\mathbf{x}) + B)$) concentrates around the optimizers of an optimization problem over $[-1,1]^n$. Analyzing this optimization problem (cf. Lemma 3.3), we show in Theorem 1.5 that for approximately regular graphs any optimizers must be nearly constant, and so $T_n(\mathbf{X}) = o_p(1)$. On the other hand, if the graph is not regular, then it is shown in Theorem 1.4 that none of the optimizers can be close to a constant vector, and so $T_n(\mathbf{X}) = \Theta_p(1)$. Thus combining Theorems 1.4 and 1.5 we have $T_n(\mathbf{X}) = o_p(1)$ iff the coupling matrix is asymptotically regular.

The mean field approach of [2] works only for graphs whose average degree goes to ∞, and it breaks down for graphs of bounded degree. We now claim that for graphs of bounded degree, we always have $T_n(\mathbf{X}) = \Theta_p(1)$ (cf. Theorem 1.8), irrespective of whether the graph is regular or not. Here, $T_n(\mathbf{X})$ is computed using the coupling matrix A_n which is now the adjacency matrix (modulo some scaling) of a bounded degree graph. The proof of Theorem 1.8 uses a change of measure argument, by showing that if $T_n(\mathbf{X}) = o_p(1)$ under the Ising model on a bounded degree graph, then the corresponding Ising model must be close to the Curie–Weiss model (via Lemmas 2.1 and 4.2), and so $T_n(\mathbf{X}) = o_p(1)$ under the Curie–Weiss model as well. But expressing the Curie–Weiss model as a mixture of i.i.d. models (owing to [19], Lemma 3), a direct argument using the bounded degree assumption shows that $T_n(\mathbf{X}) = \Theta_p(1)$ under the Curie–Weiss model (cf. Lemma 4.3), which contradicts the assumption. Thus for bounded degree graphs we always have $T_n(\mathbf{X}) = \Theta_p(1)$ irrespective of whether the graph is regular or not.

The combination of Theorems 1.4, 1.5 and 1.8 concludes that $\sqrt{n}$ consistency is always attained by the bivariate pseudo-likelihood estimator whenever the sequence of graphs is either not asymptotically regular, or the average degree diverges. To confirm that joint estimation is indeed hard on regular graphs of large degree, we consider Ising models on Erdős–Renyi graphs, and show that they are asymptotically unidentifiable along the parameter set $\Gamma_t := \{(\beta, B) \subset (0, \infty)^2 : t = \tanh(\beta t + B)\}$. To get an intuitive explanation of why this is so, suppose we want to approximate an Ising model on a regular graph by an i.i.d. model. Then one reasonable guess is to choose the mean of this i.i.d. model to be the same as $\mathbb{E}_{n,\beta_0,B_0}\bar{X}$ under the Ising model, where (β_0, B_0) is the true configuration. It follows from [2] that for $B_0 > 0$ the quantity $\mathbb{E}_{n,\beta_0,B_0}\bar{X}$ converges to t, where t is the unique positive root of the equation $t = \tanh(\beta_0 t + B_0)$. By construction, for any $(\beta, B) \in \Gamma_t$ we have $t = \tanh(\beta t + B)$, for the same t. Thus the i.i.d. approximation is the same for all $(\beta, B) \in \Gamma_t$, and so it seems reasonable that all Ising model $\mathbb{P}_{n,\beta,B}$ for $(\beta, B) \in \Gamma_t$ are asymptotically close. Theorem 1.6 confirms this heuristic for Ising model on dense Erdős–Renyi graphs.

1.4. *Simulation.* Our results demonstrate the failure of $\sqrt{n}$ consistency for the joint estimation of the parameter (β, B) when the coupling matrix A_n is a scaled adjacency matrix of a regular graph of large degree. In what follows, we address this phenomenon using simulation. Theorem 1.6 suggests that estimation for regular matrices is the hardest along the sub parameter space $\Gamma_t := \{(\beta, B) \in (0, \infty)^2 : t = \tanh(\beta t + B)\}$, where t is a fixed positive number. Setting $t = \tanh(\frac{1}{3})$, we choose two different values of the pair (β, B) in Γ_t, namely, $(\beta, B) = (0.5, 0.17)$ and $(\beta, B) = (0.8, 0.08)$. Next, we draw one random d-regular graph G_n with degree $d = 400$, with $n = 500$ nodes. For each values of (β, B), we generate 1000 samples from the Ising model with A_n being the scaled adjacency matrices for the graph G_n. On each of 2000 different samples (1000 for each (β, B)), we estimate (β, B) by solving the bivariate pseudo-likelihood equation. In Figure 2, we plot the corresponding pseudo-likelihood estimates $(\hat{\beta}_n, \hat{B}_n)$ for $(\beta, B) = (0.5, 0.17)$ and $(0.8, 0.08)$, colored in green and blue, respectively. Notice that the fit of the estimates are spread along the line of nonidentifiability. This agrees with our results, which shows that joint estimation for both parameters is harder on regular graphs of large degree.

1.5. *Main contributions and future scope.* In this paper, we provide the first rigorous results for $\sqrt{n}$ consistent estimation of both the parameters (β, B) in an Ising model, using the pseudo-likelihood estimate. Prior to this work, only estimation of the parameter β was well understood, under the assumption that $B = 0$ is known. One of the main challenges in understanding the behavior of bivariate pseudo-likelihood estimator is to understand the eigenvalues of the Hessian of the estimating equation. Theorem 1.2 takes care of this by connecting the eigenvalues of the Hessian to the statistic $T_n(\cdot)$ (see Definition 1.1), which in particular shows that if $T_n(\mathbf{X}) = \Theta_p(1)$ (does not converge to 0 in probability), the joint pseudo-likelihood estimator is $\sqrt{n}$ consistent (cf. Corollary 1.3). Understanding the behavior of the random variable $T_n(\mathbf{X})$ itself is a nontrivial task. To achieve this, we develop concentration results for the vector $(m_1(\mathbf{X}), \dots, m_n(\mathbf{X})$ for graphs of large average degree in Theorem 3.2 in the Appendix, using the mean field set up of [2]. This is the main ingredient for proving Theorems 1.4 and 1.5, which shows that $T_n(\mathbf{X}) = o_p(1)$ iff the graph sequence is dense and asymptotically regular. Since the mean field approach of [2] fails for bounded degree graphs, a different argument is needed to show $T_n(\mathbf{X}) = \Theta_p(1)$ always for bounded degree graphs, irrespective of whether the graph sequence is regular or not. This is carried out in Theorem 1.8 by a change of measure to Curie–Weiss, along with the observation that a Curie–Weiss model is a *mixture* of i.i.d. models. To confirm the intuition that the estimation is supposed to be hard on regular graphs of large degree, Theorem 1.6 gives a lower bound result, showing that consistent estimation at rate $\sqrt{n}$ is impossible by any estimator, if the underlying graph is Erdős–Renyi with parameter p fixed.

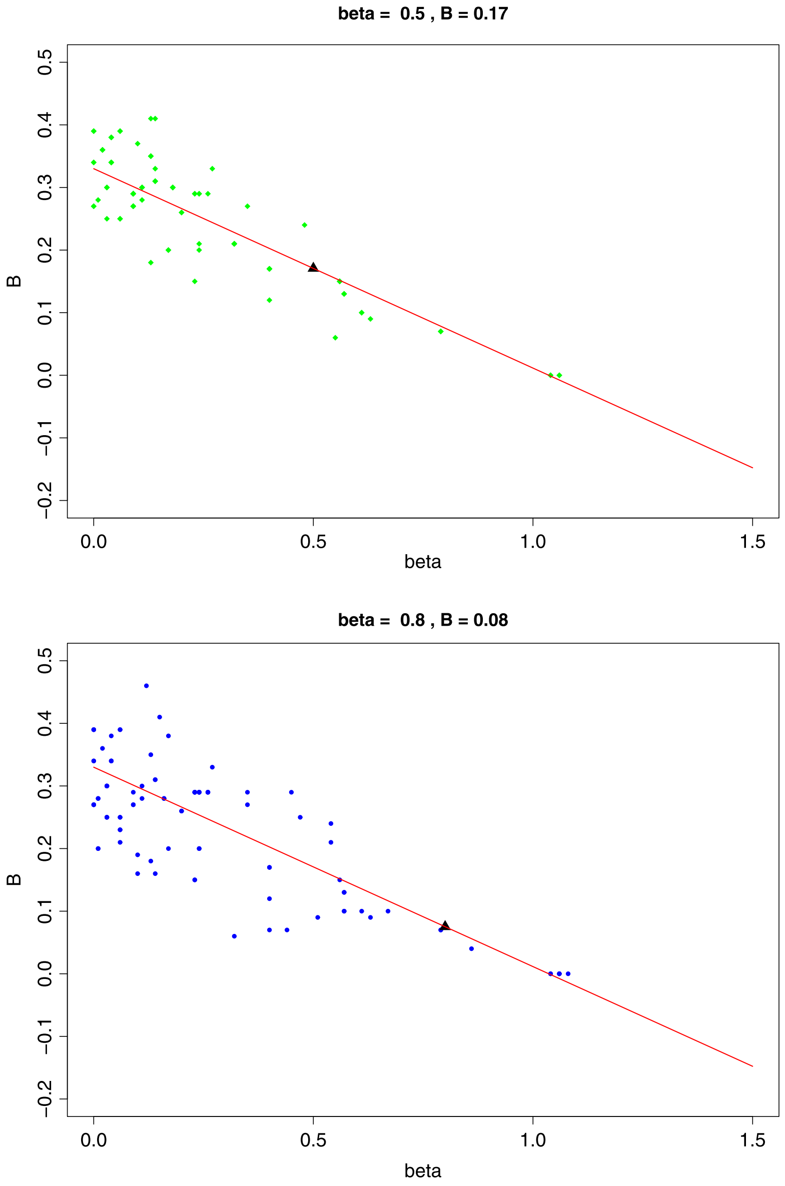

FIG. 2. *Plots of the pseudo-likelihood estimates* $(\hat{\beta}_n, \hat{B}_n)$ *for* 1000 *samples of the Ising model on a* d*-regular random graph* G *with nodes* $n = 500$ *and* $d = 400$. *We have considered the Ising model with* $(\beta, B) = (0.5, 0.17)$ *in the first figure and* $(\beta, B) = (0.8, 0.08)$ *in the second figure. In both of these two figures, the true* (β, B) *are shown as black triangle-shaped points which are lying on the line* $\tanh^{-1}(t) = t\beta + B$ (*shown in red*) *for* $t = \tanh(\frac{1}{3})$.

As part of future work, it would be interesting if one can extend Theorem 1.6 to show that consistent estimation is indeed impossible for Ising models on all asymptotically dense regular graphs, and not just Erdős–Renyi graphs. In fact, more generally we conjecture that the best rate of estimation possible on d regular graphs is $\sqrt{\frac{d}{n}}$. When the pseudo-likelihood estimator is consistent, it is natural to ask whether there is a CLT for the estimator as well. In a different direction, one can study the behavior of the MLE, and compare its performance with the pseudo-likelihood estimator. Even though it is not usually expected that the pseudo-likelihood estimator will out perform the MLE, the computational efficiency of the pseudo-likelihood estimator makes it attractive. Thus it remains to be seen whether there is a significant gain in accuracy by using the MLE. Finally, most (if not all) of our results depend crucially on the assumption that the coupling matrix is nonnegative. It is of interest to see if one can generalize our results to matrices with both positive and negative entries, such as the Sherrington–Kirkpatrick model and the Hopfield model.

The rest of the paper is outlined as follows: Section 2 details the proof of Theorem 1.2. Section 3 proves Theorem 1.4 and Theorem 1.5 with the help of Theorem 3.2, the proof of which is deferred to the Appendix. Finally, Section 4 gives the proof of Theorem 1.6 and Theorem 1.8. The proof of Proposition 1.9 is also deferred to the Appendix.

2. Proof of Theorem 1.2.

The following lemma is a collection of estimates to be used throughout the rest of this paper.

LEMMA 2.1. *Suppose* $\mathbf{X} = (X_1, \dots, X_n)$ *is an observation from the Ising model* (1.1), *where the coupling matrix* A_n *satisfies* (1.2) *and* (1.6).

Then, for any $\mathbf{x} = (x_1, \dots, x_n) \in [-1, 1]^n$, *setting*

$$f_n(\mathbf{x}) := \frac{\beta}{2}\mathbf{x}' A_n \mathbf{x} + B\sum_{i=1}^n x_i$$

and

$$b_i(\mathbf{x}) := \mathbb{E}(X_i | X_j = x_j, j \neq i) = \tanh\big(\beta m_i(\mathbf{x}) + B\big),$$

the following hold:

$$\limsup_{n\to\infty} \frac{1}{n^2}\mathbb{E}\big[f_n(\mathbf{X}) - f_n\big(b(\mathbf{X})\big)\big]^2 = 0, \tag{2.1}$$

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \big(X_i - b_i(\mathbf{X})\big)m_i(\mathbf{X})\right]^2 < \infty, \tag{2.2}$$

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \big(X_i - b_i(\mathbf{X})\big)\right]^2 < \infty. \tag{2.3}$$

PROOF OF LEMMA 2.1. Various versions of these estimates exist already in the literature. In particular, (2.1) follows on invoking [9], Lemma 3.1, or [2], Lemma 3.2, along with the assumption that A_n satisfies (1.6), and (2.2) follows on invoking [9], Lemma 3.2, along with the assumption that A_n satisfies (1.2). Finally, (2.3) follows as an easy consequence of [19], Lemma 1. □

We also need the following lemma for proving Theorem 1.2 and Proposition 1.9. The proof of the lemma is deferred to Supplement A.

LEMMA 2.2. *Suppose* $\mathbf{X} = (X_1, \dots, X_n)$ *is an observation from Ising model as in* (1.1) *such that* (1.2) *and* (1.3) *holds. If the true parameter is* $\beta_0 > 0$, $B_0 \neq 0$, *then there exists* $\delta > 0$ *such that*

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}_{n,\beta_0,B_0}\left(\left|\sum_{i=1}^n X_i m_i(\mathbf{X})\right| < n\delta\right) < 0.$$

2.1. *Proof of Theorem* 1.2.

(a) Setting

$$\widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x}) := \sum_{i=1}^n \bigl(\beta x_i m_i(\mathbf{x}) + Bx_i - \log\cosh\bigl(\beta m_i(\mathbf{x}) + B\bigr)\bigr) \tag{2.4}$$

note that $\mathrm{PL}_n(\beta, B|\mathbf{x}) = \nabla\widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x})$. Differentiating the function $(\beta, B) \mapsto \widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x})$ twice we get the negative Hessian matrix given by

$$H_n(\beta, B|\mathbf{x}) = \begin{bmatrix} \sum_{i=1}^n m_i(\mathbf{x})^2\theta_i(\beta, B|\mathbf{x}) & \sum_{i=1}^n m_i(\mathbf{x})\theta_i(\beta, B|\mathbf{x}) \\ \sum_{i=1}^n m_i(\mathbf{x})\theta_i(\beta, B|\mathbf{x}) & \sum_{i=1}^n \theta_i(\beta, B|\mathbf{x}) \end{bmatrix}, \tag{2.5}$$

where $\theta_i(\beta, B|\mathbf{x}) := \mathrm{sech}^2(\beta m_i(\mathbf{x}) + B)$. The determinant of the Hessian is given by

$$\begin{aligned} &\left[\sum_{i=1}^n m_i(\mathbf{x})^2\theta_i(\beta, B|\mathbf{x})\right] \times \left[\sum_{i=1}^n \theta_i(\beta, B|\mathbf{x})\right] - \left[\sum_{i=1}^n m_i(\mathbf{x})\theta_i(\beta, B|\mathbf{x})\right]^2 \\ &\quad = \frac{1}{2}\sum_{i,j=1}^n \theta_i(\beta, B|\mathbf{x})\theta_i(\beta, B|\mathbf{x})\bigl(m_i(\mathbf{x}) - m_j(\mathbf{x})\bigr)^2 \\ &\quad \geq \frac{1}{2}\mathrm{sech}^4\bigl(\beta\gamma + |B|\bigr)\sum_{i,j=1}^n \bigl(m_i(\mathbf{x}) - m_j(\mathbf{x})\bigr)^2 = \mathrm{sech}^4\bigl(\beta\gamma + |B|\bigr)n^2T_n(\mathbf{x}), \end{aligned}$$

which gives

$$\bigl|H_n(\beta, B|\mathbf{x})\bigr| = \lambda_n(\beta, B|\mathbf{X})\mu_n(\mathbf{x}) \geq \mathrm{sech}^4\bigl(\beta\gamma + |B|\bigr)n^2T_n(\mathbf{x}). \tag{2.6}$$

Since on $A_{1,n}^c$ we have $T_n(\mathbf{x}) > 0$ it follows that the Hessian is negative definite, and so the function $\widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x})$ is strictly concave. To show that there exists a global maximizer $(\hat{\beta}_n, \hat{B}_n)$, it thus suffices to show that

$$\lim_{\beta\to+\pm\infty} \widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x}) = -\infty, \qquad \lim_{B\to\pm\infty} \widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x}) = -\infty.$$

To see this, note that $\mathbf{x} \in A_{2,n}^c$ implies there exists $i \in [n]$ such that $x_i m_i(\mathbf{x}) = -|m_i(\mathbf{x})|$, and $m_i(\mathbf{x}) \neq 0$. Since we have

$$\widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x}) \leq \beta x_i m_i(\mathbf{x}) - \log\bigl(e^{\beta m_i(\mathbf{x})x_i + B} + e^{-\beta m_i(\mathbf{x}) + B}\bigr),$$

on letting $\beta \to \infty$ gives $\lim_{\beta\to+\infty} \widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x}) = -\infty$. A similar argument shows that if $\mathbf{x} \in A_{3,n}^c$, then $\lim_{\beta\to-\infty} \widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x}) = -\infty$. Finally, it is immediate that $\lim_{B\to\pm\infty} \widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x}) = -\infty$, for any $\mathbf{x} \notin A_{4,n}$. Thus there exists a unique global maximum $(\hat{\beta}_n, \hat{B}_n)$ for the function $(\beta, B) \mapsto \widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x}) \in \mathbb{R}^2$, and so $(\hat{\beta}_n, \hat{B}_n)$ is the unique root of $\mathrm{PL}_n(\beta, B|\mathbf{x})$.

We will now show that if $\mathbf{x} \in A_{j,n}$ for some $j = 1, 2, 3, 4$, then the pseudo-likelihood estimator is not defined.

- $\mathbf{x} \in A_{1,n}$

  On this set, we have $T_n(\mathbf{x}) = 0$ which implies $m_i(\mathbf{x}) = \bar{m}(\mathbf{x})$ for all $i \in [n]$. This implies that $Q_n(\beta, B|\mathbf{x}) = \bar{m}(\mathbf{x})R_n(\beta, B|\mathbf{x})$, and so the equation $\mathrm{PL}_n(\beta, B|\mathbf{x}) = (0, 0)$ is equivalent to

$$R_n(\beta, B|\mathbf{x}) = 0 \quad \Leftrightarrow \quad \bar{\mathbf{x}} = \tanh\bigl(\beta\bar{m}(\mathbf{x}) + B\bigr).$$

Since the function $(\beta, B) \mapsto \widetilde{\mathrm{PL}}_n(\beta, B|\mathbf{x})$ is convex, it follows that any (β, B) satisfying this equation is a global maximizer, and hence in this case the set of maximizers is a line in the two-dimensional plane, and hence not unique. Thus the pseudo-likelihood estimator is not defined.

- $\mathbf{x} \in A_{2,n}$

  On this set, we have

$$Q_n(\beta, B|\mathbf{x}) = \sum_{i=1}^n |m_i(\mathbf{x})| - \sum_{i=1}^n m_i(\mathbf{x}) \tanh\bigl(\beta m_i(\mathbf{x}) + B\bigr) > 0,$$

  and so the equation $Q_n(\beta, B|\mathbf{x}) = 0$ has no roots in $\mathbb{R}^2$, and so the pseudo-likelihood estimator is not defined.

- $\mathbf{x} \in A_{3,n}$

  Similarly, on this set $Q_n(\beta, B|\mathbf{x}) < 0$ for all $(\beta, B) \in \mathbb{R}^2$, and so the pseudo-likelihood estimator is not defined.

- $\mathbf{x} \in A_{4,n}$

  If $\mathbf{x} = \mathbf{1}$, then we have

$$R_n(\beta, B|\mathbf{x}) = \sum_{i=1}^n \bigl(1 - \tanh\bigl(\beta m_i(\mathbf{x}) + B\bigr)\bigr) > 0,$$

  and so the equation $R_n(\beta, B|\mathbf{x}) = 0$ has no roots in $\mathbb{R}^2$, and so the pseudo-likelihood estimator is not defined.

  Similarly, if $\mathbf{x} = -\mathbf{1}$, then $R_n(\beta, B|\mathbf{x}) < 0$ for all $(\beta, B) \in \mathbb{R}^2$.

(b) Note that if $\mathbf{x} \in A_{2,n}$ we have

$$Q_n(\beta_0, B_0|\mathbf{x}) \geq (1 - \tanh\bigl(\beta_0\gamma + |B_0|\bigr) \sum_{i=1}^n |m_i(\mathbf{x})|,$$

which gives

$$\left|\sum_{i=1}^n x_i m_i(\mathbf{x})\right| \leq \sum_{i=1}^n |m_i(\mathbf{x})| \leq \frac{1}{1 - \tanh(\beta_0\gamma + |B_0|)} Q_n(\beta_0, B_0|\mathbf{x}).$$

Since $Q_n(\beta_0, B_0|\mathbf{X}) = O_p(\sqrt{n})$ by (2.2) and $\sum_{i=1}^n X_i m_i(\mathbf{X})$ is not $o_p(n)$ by Lemma 2.2, $\mathbb{P}_{n,\beta_0,B_0}(A_{2,n})$ converges to 0. A similar proof takes care of $A_{3,n}$. It thus remains to show that $\mathbb{P}_{n,\beta_0,B_0}(A_{4,n})$ converges to 0 as well. To this effect, note that if $\mathbf{x} = \pm\mathbf{1}$, then we have $|R_n(\beta_0, B_0|\mathbf{x})| \geq n(1 - \tanh(\beta_0\gamma + |B_0|)$, the probability of which converges to 0 as $R_n(\beta_0, B_0|\mathbf{X}) = O_p(\sqrt{n})$ by (2.3).

(c) By part (b), we have $\mathbf{x} \in A_{2,n}^c \cap A_{3,n}^c \cap A_{4,n}^c$ with probability tending to 1. Also by assumption, we have $T_n(\mathbf{X}) > 0$ with probability tending to 1, and so the pseudo-likelihood estimator $(\hat{\beta}_n, \hat{B}_n)$ is well defined with probability tending to 1.

Recall the 2×2 matrix $H_n(\beta, B|\mathbf{x})$ as defined in (2.5), and denote $\lambda_n(\beta, B|\mathbf{x}) \geq \mu_n(\beta, B|\mathbf{x})$ to be its eigenvalues. We start by giving a lower bound to the minimum eigenvalue $\mu_n(\beta, B|\mathbf{x})$. To this effect, note that

$$\begin{aligned}\lambda_n(\beta, B|\mathbf{x}) + \mu_n(\beta, B|\mathbf{x}) = \mathrm{tr}(H_n(\beta, B|\mathbf{x}) &= \sum_{i=1}^n \theta_i(\beta, B|\mathbf{x})\bigl(m_i^2(\mathbf{x}) + 1\bigr) \\ &\leq n\bigl(1 + \gamma^2\bigr),\end{aligned}$$

which along with (2.6) gives

$$\tag{2.7} \begin{aligned} \mu_n(\beta, B|\mathbf{x}) &\geq \frac{\lambda_n(\beta, B|\mathbf{x})\mu_n(\beta, B|\mathbf{x})}{\lambda_n(\beta, B|\mathbf{x}) + \mu_n(\beta, B|\mathbf{x})} = \frac{|H_n(\beta, B|\mathbf{x})|}{\operatorname{tr}(H_n(\beta, B|\mathbf{x})} \\ &\geq \frac{\operatorname{sech}^4(\beta\gamma + |B|)}{1+\gamma^2} nT_n(\mathbf{x}). \end{aligned}$$

Armed with this estimate, we now complete the proof of the Theorem. To this effect, setting $(\beta_t, B_t) = (t\hat{\beta}_n + (1-t)\beta_0, t\hat{B}_n + (1-t)B_0)$, define a function $g_n : [0, 1] \to \mathbb{R}$ by

$$g_n(t) := (\hat{\beta}_n - \beta_0)Q_n(\beta_t, B_t|\mathbf{x}) + (\hat{B}_n - B_0)R_n(\beta_t, B_t|\mathbf{x}),$$

and note that

$$\tag{2.8} \begin{aligned} \left|g_n(1) - g_n(0)\right| &= \left|(\hat{\beta}_n - \beta_0)Q_n(\beta_0, B_0|\mathbf{x}) + (\hat{B}_n - B_0)R_n(\beta_0, B_0|\mathbf{x})\right| \\ &= O_p(\sqrt{n}Y_n), \end{aligned}$$

where $Y_n := \|\hat{\beta}_n - \beta_0, \hat{B}_n - B_0\|_2$, and we use the Cauchy–Schwarz inequality along with (2.2) and (2.3) of Lemma 2.1. Also we have

$$g_n'(t) = (\hat{\beta}_n - \beta_0, \hat{B}_n - B_0)H_n(\beta_t, B_t|\mathbf{x})(\hat{\beta}_n - \beta_0, \hat{B}_n - B_0)^\top \geq \mu_n(\beta_t, B_t|\mathbf{x})Y_n^2.$$

In particular, we have $g_n'(t) \geq 0$ for all $t \in (0, 1)$. Further, using (2.7) we get the existence of $r, s > 0$ such that

$$\inf_{\beta > -, B \neq 0, \|\beta - \beta_0, B - B_0\| \leq r} \mu_n(\beta, B|\mathbf{x}) \geq snT_n(\mathbf{x}).$$

Noting that $\|\beta_t - \beta_0, B_t - B_0\|_2 = tY_n$ gives

$$\tag{2.9} \int_0^1 g_n'(t)\,dt \geq \int_0^{\min(1, \frac{r}{Y_n})} g_n'(t)\,dt \geq \min\left(1, \frac{r}{Y_n}\right) snT_n(\mathbf{x})W_n^2,$$

which along with (2.8) gives $\min(Y_n, r) = O_p(\frac{1}{\sqrt{n}T_n(\mathbf{X}})$. Since $r > 0$ is fixed, it follows that $Y_n = o_p(1)$, and so $(\hat{\beta}_n, \hat{B}_n)$ converges in probability to (β_0, B_0). This shows that $Y_n < r$ with probability tending to 1, which on using (2.9) gives $\int_0^1 g_n'(t)\,dt \geq snT_n(\mathbf{x})Y_n^2$. Along with (2.8), this gives $nT_n(\mathbf{X})Y_n = O_p(\sqrt{n})$, which is the claimed bound.

3. Proofs of Theorem 1.4 and Theorem 1.5. The main tool required for proving Theorem 1.4 and Theorem 1.5 is the following theorem, which proves a large deviation estimate for Ising models that might be of independent interest. The proof of this theorem is very similar to the proof of [9], Theorem 1.6, and [2], Theorem 1.1, and is placed in the Appendix. See also [13], Corollary 12, which proves a similar result for Ising models under identical conditions (i.e., (1.2) and (1.6)). The main difference is that [13] expresses mean field Ising models as mixture of i.i.d. laws, whereas we focus on the behavior of $m(\mathbf{X})$ which is more relevant to the criterion $T_n(\mathbf{X}) = \Theta_p(1)$ provided in Corollary 1.3.

DEFINITION 3.1. For any $\mathbf{y} \in [-1, 1]^n$, define a vector $b(\mathbf{y}) \in [-1, 1]^n$ by setting $b_i(\mathbf{y}) := \tanh(\beta m_i(\mathbf{y}) + B) \in [-1, 1]$.

The next result gives a crucial large deviation estimate for the random vector $b(\mathbf{X})$. We defer its proof to Supplement A.

THEOREM 3.2. *Suppose* $\mathbf{X} = (X_1, \dots, X_n)$ *is an observation from the Ising model* (1.1), *where the coupling matrix* A_n *satisfies* (1.2) *and* (1.6).

(a) *Let* $r_n := \sup_{\mathbf{y}\in[-1,1]^n}\{f_n(\mathbf{y}) - I(\mathbf{y})\}$ *where*

$$f_n(\mathbf{y}) := \frac{\beta}{2}\mathbf{y}'A_n\mathbf{y} + B\sum_{i=1}^n y_i,$$

$$I(\mathbf{y}) := \sum_{i=1}^n \left\{\frac{1+y_i}{2}\log\frac{1+y_i}{2} + \frac{1-y_i}{2}\log\frac{1-y_i}{2}\right\}.$$

Then we have

$$f_n\big(b(\mathbf{X})\big) - I\big(b(\mathbf{X})\big) - r_n = o_p(n).$$

(b) *Further, we have*

$$\big\|\nabla f_n\big(b(\mathbf{X})\big) - \nabla I\big(b(\mathbf{X})\big)\big\| = o_p(\sqrt{n}).$$

3.1. *Proof of Theorem* 1.4. Since $T_n = O_p(1)$, it suffices to show that $\frac{1}{T_n} = O_p(1)$, which is equivalent to showing that for any sequence $\{\varepsilon_n\}_{n\geq 1}$ converging to 0 we have

$$\lim_{n\to\infty}\mathbb{P}_{n,\beta,B}\left(\sum_{i=1}^n\big(m_i(\mathbf{X}) - \bar{m}(\mathbf{X})\big)^2 \leq n\varepsilon_n\right) = 0. \tag{3.1}$$

Since the set of probability measures on $[-\gamma,\gamma]$ is compact with respect to weak topology, without loss of generality by passing to subsequence we can assume the sequence of empirical measures $\frac{1}{n}\sum_{i=1}^n \delta_{\mathcal{R}_n(i)}$ converge weakly to μ, where μ is a probability measure on $[-\gamma,\gamma]$. This along with the dominated convergence theorem and (1.7) gives

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\big(\mathcal{R}_n(i) - \bar{\mathcal{R}}_n\big)^2 = \int_{[-\gamma,\gamma]}(\theta - \mathbb{E}_\mu\theta)^2\,d\mu(\theta) > 0. \tag{3.2}$$

Proceeding to show (3.1), first note that

$$\max_{i\in[n]}\big|b_i(\mathbf{x})\big| = \max_{i\in[n]}\big|\tanh\big(\beta m_i(\mathbf{x}) + B\big)\big| \leq \tanh\big(\beta\gamma + |B|\big) =: p < 1, \tag{3.3}$$

and so $b_i(\mathbf{x}) \in [-p,p]$. Also use (1.2) to note that there exists a finite positive constant $C(\beta,B,\gamma)$ such that

$$\begin{aligned}\sum_{i=1}^n\big(m_i(\mathbf{x}) - \bar{m}(\mathbf{x})\big)^2 &= \frac{1}{2n}\sum_{i,j=1}^n\big(m_i(\mathbf{x}) - m_j(\mathbf{x})\big)^2\\ &\geq \frac{C(\beta,B,\gamma)}{2n}\sum_{i,j=1}^n\big(b_i(\mathbf{x}) - b_j(\mathbf{x})\big)^2\\ &= C(\beta,B,\gamma)\sum_{i=1}^n\big(b_i(\mathbf{x}) - \bar{b}(\mathbf{x})\big)^2,\end{aligned}$$

and so changing variables to $\delta_n := \varepsilon_n/C(\beta,B,\gamma)$ for verifying (3.1) it suffices to check that

$$\lim_{n\to\infty}\mathbb{P}_{n,\beta,B}\left(\sum_{i=1}^n\big(b_i(\mathbf{X}) - \bar{b}(\mathbf{X})\big)^2 \leq n\delta_n, \max_{i\in[n]}\big|b_i(\mathbf{X})\big| \leq p\right) = 0. \tag{3.4}$$

To show this, note that

$$\nabla f_n(\mathbf{y}) - \nabla I(\mathbf{y}) = \beta A_n\mathbf{y} + B\mathbf{1} - \operatorname{arctanh}(\mathbf{y}).$$

Thus if $\mathbf{y}$ is such that $\sum_{i=1}^n (y_i - \bar{y})^2 \le n\delta_n$ and $\max_{i\in[n]} |y_i| \le p$, then with $\tilde{\mathbf{y}} := \bar{y}\mathbf{1}$ Triangle inequality gives

$$\begin{aligned}\|\nabla f_n(\mathbf{y}) - \nabla I(\mathbf{y})\| &\ge \|\nabla f_n(\tilde{\mathbf{y}}) - \nabla I(\tilde{\mathbf{y}})\| - \|\mathbf{y} - \tilde{\mathbf{y}}\|\Big(\beta\|A_n\|_2 + \frac{1}{1-p^2}\Big) \\ &\ge \|\nabla f_n(\tilde{\mathbf{y}}) - \nabla I(\tilde{\mathbf{y}})\| - \sqrt{n\delta_n}\Big(\beta\gamma + \frac{1}{1-p^2}\Big) \\ &= \|\nabla f_n(\tilde{\mathbf{y}}) - \nabla I(\tilde{\mathbf{y}})\| - o(\sqrt{n}).\end{aligned}$$

Finally, we have

$$\begin{aligned}\|\nabla f_n(\tilde{\mathbf{y}}) - \nabla I(\tilde{\mathbf{y}})\|^2 &= \sum_{i=1}^n \big(\beta\bar{y}\mathcal{R}_n(i) + B - \operatorname{arctanh}(\bar{y})\big)^2 \\ &\ge \inf_{t\in[-p,p]} \sum_{i=1}^n \big(\beta t\mathcal{R}_n(i) + B - \operatorname{arctanh}(t)\big)^2 \\ &= n \inf_{t\in[-p,p]} \int_{-\gamma}^{\gamma} \big(\beta t\theta + B - \operatorname{arctanh}(t)\big)^2\, d\mu(\theta) + o(n).\end{aligned}$$

Combining these estimates, on the set $\sum_{i=1}^n (b_i(\mathbf{x}) - \bar{b}(\mathbf{x}))^2 \le n\delta_n$ we have

$$\frac{1}{\sqrt{n}}\big\| f_n\big(b(\mathbf{x})\big) - I\big(b(\mathbf{x})\big)\big\| \ge \sqrt{\inf_{t\in[-p,p]} \int_{-\gamma}^{\gamma} \big(\beta t\theta + B - \operatorname{arctanh}(t)\big)^2\, d\mu(\theta)} - o(1),$$

from which the desired conclusion follows via part (b) of Lemma 3.2, if we can show that

$$\inf_{t\in[-p,p]} \int_{-\gamma}^{\gamma} \big(\beta t\theta + B - \operatorname{arctanh}(t)\big)^2\, d\mu(\theta) > 0.$$

If not, then there exists $t \in [-p, p]$ such that $\int_{-\gamma}^{\gamma} (\beta t\theta + B - \operatorname{arctanh}(t))^2\, d\mu(\theta) = 0$. If $t = 0$, then we have

$$0 = \int_{-\gamma}^{\gamma} \big(\beta t\theta + B - \operatorname{arctanh}(t)\big)^2\, d\mu(\theta) = B^2 \int_{-\gamma}^{\gamma} \mu(d\theta) = B^2 \ne 0,$$

a contradiction. Finally, if $t \ne 0$, then we have $\theta \overset{a.s.}{=} \frac{\operatorname{arctanh}(t) - B}{\beta t}$ is a degenerate random variable, a contradiction to (3.2). This completes the proof of the theorem.

3.2. *Proof of Theorem* 1.5. For proving Theorem 1.5, we need the following lemma, the proof of which follows by simple analysis and can be found, for example, in [12], page 10.

LEMMA 3.3. *Fix $\beta > 0$, $B \ne 0$, and define the function*

$$\phi(y) := \frac{\beta}{2}y^2 + By - I(y), \quad y \in [-1, 1].$$

Then the function $\phi(\cdot)$ has a unique global maximum at some $m_0 \in (-1, 1)$.

PROOF OF THEOREM 1.5. Fixing $\varepsilon > 0$, it suffices to show that

$$\lim_{n\to\infty} \mathbb{P}_{n,\beta,B}\Bigg(\sum_{i=1}^n \big(m_i(\mathbf{x}) - \bar{m}(\mathbf{x})\big)^2 > n\varepsilon\Bigg) = 0.$$

To this effect, note that $\bar{\mathcal{R}}_n$ is a bounded sequence of real numbers by (1.2), and so without loss of generality by passing to a subsequence, we can assume that $\bar{\mathcal{R}}_n$ converges to θ, say.

Note that this also gives

$$\sum_{i=1}^{n} \big(\mathcal{R}_n(i)-\theta\big)^2 = o(n). \tag{3.5}$$

Now, use (1.2) to note that there exists a finite positive constant $C(\beta, B, \gamma)$ such that

$$\begin{aligned}\sum_{i=1}^{n}\big(m_i(\mathbf{x})-\bar{m}(\mathbf{x})\big)^2 &= \frac{1}{2n}\sum_{i,j=1}^{n}\big(m_i(\mathbf{x})-m_j(\mathbf{x})\big)^2\\ &\le \frac{C(\beta, B,\gamma)}{2n}\sum_{i,j=1}^{n}\big(b_i(\mathbf{x})-b_j(\mathbf{x})\big)^2\\ &= C(\beta, B,\gamma)\sum_{i=1}^{n}\big(b_i(\mathbf{x})-\bar{b}(\mathbf{x})\big)^2,\end{aligned}$$

and so with $\delta := \varepsilon/C(\beta, B, \gamma)$ it suffices to check that

$$\lim_{n\to\infty}\mathbb{P}_{n,\beta,B}\left(\sum_{i=1}^{n}\big(b_i(\mathbf{x})-\bar{b}(\mathbf{x})\big)^2 > n\delta\right) = 0.$$

Let $\mathbf{y}\in[-1,1]^n$ be any vector such that $\sum_{i=1}^n (y_i-\bar{y})^2 \ge n\delta$, and define a matrix $A_n^{(t)}$ by setting

$$A_n^{(t)}(i,j) := \begin{cases} A_n(i,j) & \text{if } \max\big(|\mathcal{R}_n(i)-\theta|, |\mathcal{R}_n(j)-\theta|\big)\le t,\\ 0 & \text{otherwise.}\end{cases}$$

Then we have

$$\begin{aligned}\sum_{i,j=1}^{n} A_n(i,j)y_iy_j &\le \sum_{i,j=1}^{n} A_n^{(t)}(i,j)y_iy_j + 2\sum_{i:|\mathcal{R}_n(i)-\theta|>t}\sum_{j=1}^{n} A_n(i,j)y_iy_j\\ &\le \sum_{i,j=1}^{n} A_n^{(t)}(i,j)y_iy_j + 2\gamma\big|\{i\in[n]:|\mathcal{R}_n(i)-\theta|>t\}\big| \qquad (3.6)\\ &= \sum_{i,j=1}^{n} A_n^{(t)}(i,j)y_iy_j + o(n),\end{aligned}$$

where the last equality uses (3.5). Since $\sum_{j=1}^n A_n^{(t)}(i,j)\le\theta+t$, it follows all eigenvalues of $A_n^{(t)}$ are bounded above by $\theta+t$, and so

$$\sum_{i,j=1}^{n} A_n^{(t)}y_iy_j = \mathbf{y}'A_n^{(t)}\mathbf{y}\le(\theta+t)\sum_{i=1}^{n}y_i^2,$$

which along with (3.6) gives

$$\begin{aligned}f_n(\mathbf{y})-I(\mathbf{y}) &= \frac{\beta}{2}\mathbf{y}'A_n\mathbf{y}+B\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}I(y_i)\\ &\le \frac{\beta}{2}(\theta+t)\sum_{i=1}^{n}y_i^2+B\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}I(y_i)+o(n) \qquad (3.7)\\ &\le n\beta t+\sum_{i=1}^{n}\phi(y_i)+o(n),\end{aligned}$$

where $\phi(y) := \frac{\beta\theta}{2}y^2 + By - I(y)$. By Lemma 3.3, it follows that $\phi(\cdot)$ has a unique global maximum in $[-1, 1]$ at some point $m_0 \in (-1, 1)$. Define another function $\Phi : [-1, 1] \mapsto [0, \infty)$ by setting

$$\Phi(y) := \frac{\phi(m) - \phi(y)}{(y - m_0)^2} \quad \text{for } y \neq m_0,$$

and note that $\Phi(\cdot)$ is strictly positive for all $y \in [-1, 1]$ other than m_0 and satisfies

$$\lim_{y \to m_0} \Phi(y) = -2\phi''(m_0) > 0.$$

Consequently, $\Phi(y)$ extends to a strictly positive continuous function on $[-1, 1]$, and so $\alpha := \inf_{y \in [-1,1]} \Phi(y) > 0$, which in turn implies

$$\phi(y) \le \phi(m_0) - \alpha(y - m_0)^2$$

for all $y \in [-1, 1]$. This, along with (3.7) gives

$$\sup_{\mathbf{y}:\sum_{i=1}^n (y_i - \bar{y})^2 > n\delta} \{f_n(\mathbf{y}) - I(\mathbf{y})\} \le o(n) + n\beta t + n\phi(m_0) - n\alpha\delta, \tag{3.8}$$

where the last inequality also uses the fact that

$$\sum_{i=1}^n (y_i - m_0)^2 \ge \sum_{i=1}^n (y_i - \bar{y})^2 \ge n\delta.$$

To complete the proof, restricting the sup over all vector $\mathbf{y}$ which are constant we get

$$\begin{aligned}\sup_{\mathbf{y} \in [-1,1]^n} \{f_n(\mathbf{y}) - I(\mathbf{y})\} &\ge n \sup_{y \in [-1,1]} \left\{\frac{\beta}{2n} y^2 \mathbf{1}' A_n \mathbf{1} + By - I(y)\right\} \\ &= o(n) + n \sup_{y \in [-1,1]} \left\{\frac{\beta}{2}\theta y^2 + By - I(y)\right\} \qquad (3.9) \\ &= o(n) + n\phi(m_0),\end{aligned}$$

where the intermediate step uses (3.5) to note that

$$\mathbf{1}' A_n \mathbf{1} = \sum_{i=1}^n \mathcal{R}_n(i) = n\theta + \sum_{i=1}^n (\mathcal{R}_n(i) - \theta) = n\theta + o(n).$$

Thus combining (3.8) and (3.9) gives

$$\sup_{\mathbf{y} \in [-1,1]^n} \{f_n(\mathbf{y}) - I(\mathbf{y})\} - \sup_{\mathbf{y}:\sum_{i=1}^n (y_i - \bar{y})^2 > n\delta} \{f_n(\mathbf{y}) - I(\mathbf{y})\} \ge n\alpha\delta - n\beta t + o(n),$$

from which the desired conclusion follows on using Theorem 3.2, since $t > 0$ is arbitrary. □

4. Proof of Theorem 1.6 and Theorem 1.8.

4.1. *Proof of Theorem* 1.6. To see why the fact that $\mathbb{Q}_n \times \mathcal{G}(n, p)$ is contiguous to $\mathbb{P}^{\text{er}}_{n,\beta,B}$ implies nonexistence of consistent estimators, suppose there exists a consistent estimator $(\tilde{\beta}_n, \tilde{B}_n)$ on Γ_t. Now fixing (β_1, B_1) and (β_2, B_2) in Γ_t, there exists disjoint open balls $\mathcal{B}_1$, $\mathcal{B}_2$ in $\mathbb{R}^2$ such that $(\beta_i, B_i) \in \mathcal{B}_i$ for $i = 1, 2$. Consistency implies

$$\lim_{n \to \infty} \mathbb{P}^{\text{er}}_{n,\beta_i,B_i}((\tilde{\beta}_n, \tilde{B}_n) \in \mathcal{B}_i) = 1,$$

which along with contiguity gives

$$\lim_{n\to\infty}\big(\mathbb{Q}_n \times \mathcal{G}(n,p)\big)\big((\tilde{\beta}_n, \tilde{B}_n) \in \mathcal{B}_i)\big) = 1.$$

But this is a contradiction, as $\mathcal{B}_1$ and $\mathcal{B}_2$ are disjoint, thus completing the proof of the theorem.

The rest of the proof is broken into two parts: Part (a) shows that the probability sequence $\mathbb{Q}_n$ is contiguous to the Curie–Weiss model $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}$ (Ising model on the complete graph); Part (b) then shows that $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B} \times \mathcal{G}(n,p)$ is contiguous to $\mathbb{P}^{\mathrm{er}}_{n,\beta,B}$.

(a) By [19], Lemma 3, we have the existence of a random variable W_n with density proportional to $e^{-nf(w)}$, where $f(w) := \beta w^2/2 + Bw - \log\cosh(w)$. Also given $W_n = w$ we have $X_1, \dots, X_n$ are i.i.d. random variables on $\{-1, 1\}$ such that

$$\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(X_i = 1 | W_n = w) = \frac{e^{\beta w + B}}{e^{\beta w + B} + e^{-\beta w - B}} = 1 - \mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(X_i = -1).$$

Finally, the conditional distribution of W_n given $X_1 = x_1, \dots, X_n = x_n$ is $N(\bar{x}, \frac{1}{n\beta})$. By a slight abuse of notation, we use $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}$ to denote the joint law of $(X_1, \dots, X_n)$ and W_n on $\{-1, 1\}^n \times \mathbb{R}$. Similarly, extend $\mathbb{Q}_n$ to $\{-1, 1\}^n \times \mathbb{R}$ by setting W_n to be independent of $(X_1, \dots, X_n)$ with a density proportional to $e^{-nf(w)}$. Thus under both $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}$ and $\mathbb{Q}_n$ the marginal distribution of W_n is the same.

We now show that $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}$ is contiguous to $\mathbb{Q}_n$. To this effect, using [14] under $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}$ we have

$$\sqrt{n}(\bar{X}_n - t) \xrightarrow{d} N\left(0, \frac{1-t^2}{1-\beta(1-t^2)}\right).$$

This implies

$$\sqrt{n}(W_n - t) \xrightarrow{d} N\left(0, \frac{1}{\beta[1-\beta(1-t^2)]}\right)$$

under both $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}$ and $\mathbb{Q}_n$, as the marginal law of W_n is the same under both measures. Using this along with a one term Taylor's expansion gives

$$\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(X_i = 1 | W_n) = \frac{1}{1 + e^{-2\beta W_n - 2B}} = \frac{1}{1 + e^{-2\beta t - 2B}} + \frac{\widetilde{W}_n}{\sqrt{n}} = \alpha + \frac{\widetilde{W}_n}{\sqrt{n}},$$

where $\widetilde{W}_n := \xi_n \sqrt{n}(W_n - t)$ for some bounded random variable ξ_n, and $\alpha := \frac{1}{1+e^{-2\beta t - 2B}}$. Since $\sqrt{n}(W_n - t)$ is $O_p(1)$ under $\mathbb{Q}_n$, it follows that $\widetilde{W}_n = O_p(1)$ as well. Also, setting $S_n := |i \in [n] : X_i = 1|$ we have $\frac{S_n - n\alpha}{\sqrt{n}} = O_p(1)$ under $\mathbb{Q}_n$. On the set $|\widetilde{W}_n| \le K$ and $|S_n - n\alpha| \le K\sqrt{n}$, we have

$$\begin{aligned}
&\log \frac{\mathbb{Q}_n(X_1 = x_1, \dots, X_n = x_n | W_n = w)}{\mathbb{P}^{\mathrm{cw}}_{n,\beta_1,B_1}(X_1 = x_1, \dots, X_n = x_n | W_n = w)} \\
&= -S_n \log \frac{\alpha + \frac{\widetilde{W}_n}{\sqrt{n}}}{\alpha} - (n - S_n) \log \frac{1 - \alpha - \frac{\widetilde{W}_n}{\sqrt{n}}}{1-\alpha} \\
&= -S_n\left[\frac{\widetilde{W}_n}{\alpha\sqrt{n}} + O\left(\frac{K^2}{n}\right)\right] + (n - S_n)\left[\frac{\widetilde{W}_n}{(1-\alpha)\sqrt{n}} + O\left(\frac{K^2}{n}\right)\right] \\
&= -\frac{\widetilde{W}_n}{\sqrt{n}}\left[\frac{S_n}{\alpha} - \frac{n - S_n}{1-\alpha}\right] + O(K^2)
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{\widetilde{W}_n}{\sqrt{n}}\left[\frac{S_n - n\alpha}{\alpha(1-\alpha)}\right] + O(K^2) \\
&\le \frac{K^2}{\alpha(1-\alpha)} + O(K^2) =: z_K.
\end{aligned}
$$

Thus if $A_n \subset \{-1,1\}^n$ is any sequence of sets such that $\lim_{n\to\infty} \mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(\mathbf{X}\in A_n) = 0$, then denoting $C_n := (\int_{\mathbb{R}} e^{-f_n(w)}\,dw)^{-1}$ we have

$$
\begin{aligned}
&\mathbb{Q}_n\big(\mathbf{X}\in A_n, |\widetilde{W}_n| \le K, |S_n - np| \le K\sqrt{n}\big) \\
&\quad = C_n \int_{\mathbb{R}} \mathbb{Q}_n\big(\mathbf{X}\in A_n, |\widetilde{W}_n| \le K, |S_n - np| \le K\sqrt{n} | W_n = w\big) e^{-f_n(w)}\,dw \\
&\quad \le C_n z_K \int_{\mathbb{R}} \mathbb{P}^{\mathrm{cw}}_{n,\beta,B}\big(A_n, |\widetilde{W}_n| \le K, |S_n - np| \le K\sqrt{n} | W_n = w\big) e^{-f_n(w)}\,dw \\
&\quad = z_K \mathbb{P}^{\mathrm{cw}}_{n,\beta,B}\big(\mathbf{X}\in A_n, |\widetilde{W}_n| \le K, |S_n - np| \le K\sqrt{n}\big) \\
&\quad \le z_K P^{\mathrm{cw}}_{n,\beta,B}(\mathbf{X}\in A_n).
\end{aligned}
$$

This gives

$$
\mathbb{Q}_n(\mathbf{X}\in A_n) \le z_K \mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(\mathbf{X}\in A_n) + \mathbb{Q}_n\big(|\widetilde{W}_n| > K\big) + \mathbb{Q}_n\big(|S_n - np| > K\sqrt{n}\big),
$$

which on letting $n\to\infty$ followed by $K\to\infty$ gives

$$
\limsup_{n\to\infty} \mathbb{Q}_n(\mathbf{X}\in A_n) = 0,
$$

and so $\mathbb{Q}_n$ is contiguous to $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}$ and the proof is complete. Even though we do not need it, we note that in this case a symmetric proof gives the reverse conclusion as well, that is, $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}$ and $\mathbb{Q}_n$ are mutually contiguous.

(b) We now show that $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B} \times \mathcal{G}(n,p)$ is contiguous to $\mathbb{P}^{\mathrm{er}}_{n,\beta,B}$, for which invoking Proposition 6.1 of [5] it further suffices to show that $D(\mathbb{P}^{\mathrm{cw}}_{n,\beta,B} \times \mathcal{G}(n,p)||\mathbb{P}^{\mathrm{er}}_{n,\beta,B}) = O(1)$, where $D(\cdot||\cdot)$ is the Kullback–Leibler divergence. A direct computation shows that $D(\mathbb{P}^{\mathrm{cw}}_{n,\beta,B} \times \mathcal{G}(n,p)||\mathbb{P}^{\mathrm{er}}_{n,\beta,B})$ equals

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{G}(n,p)} \sum_{\mathbf{x}\in\{-1,1\}^n} \mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(\mathbf{x}) \left(\frac{\beta}{n-1}\sum_{i,j=1}^n \left[1 - \frac{G_n(i,j)}{p}\right] x_i x_j + \log\frac{Z_n^{\mathrm{er}}(\beta,B)}{Z_n^{\mathrm{cw}}(\beta,B)}\right) \\
&\quad = \mathbb{E}_{\mathcal{G}(n,p)} \log Z_n^{\mathrm{er}}(\beta,B) - \log Z_n^{\mathrm{cw}}(\beta,B) \\
&\quad \le \log \mathbb{E}_{\mathcal{G}(n,p)} Z_n^{\mathrm{er}}(\beta,B) - \log Z_n^{\mathrm{cw}}(\beta,B),
\end{aligned}
$$

where the last inequality is by Jensen's inequality, and $Z_n^{\mathrm{cw}}(\beta,B)$ and $Z_n^{\mathrm{er}}(\beta,B)$ denote the normalizing constants for the corresponding Ising models. Finally, note that

$$
\begin{aligned}
\mathbb{E}_{\mathcal{G}(n,p)} Z_n^{\mathrm{er}}(\beta,B) &= \sum_{\mathbf{x}\in\{-1,1\}^n} \prod_{1\le i<j\le n}^{n} \mathbb{E}_{\mathcal{G}(n,p)} e^{\frac{\beta}{(n-1)p} G_n(i,j) x_i x_j + B\sum_{i=1}^n x_i} \\
&\le \sum_{\mathbf{x}\in\{-1,1\}^n} \prod_{1\le i<j\le n} \exp\left(\frac{\beta}{n-1} x_i x_j + \frac{\beta^2}{8(n-1)^2 p^2}\right) e^{B\sum_{i=1}^n x_i},
\end{aligned}
$$

where we use Hoeffding's lemma to get $\mathbb{E}e^{t(\mathrm{Bin}(1,p)-p)} \le e^{\frac{t^2}{8}}$ for $t \in \mathbb{R}$ and $p \in (0,1)$. Combining we have

$$\begin{aligned}\mathbb{E}_{\mathcal{G}(n,p)} Z_n^{\mathrm{er}}(\beta,B) &\le e^{\frac{\beta^2}{8p^2}} \sum_{\vec{x}\in\{-1,1\}^n} \exp\left\{\frac{\beta}{n-1}\sum_{1\le i<j\le n} x_i x_j + B\sum_{i=1}^n x_i\right\}\\ &= e^{\frac{\beta^2}{8p^2}} Z_n^{\mathrm{cw}}(\beta,B),\end{aligned}$$

which gives an upper bound to the Kullback–Leibler divergence which is independent of n, and hence completes the proof of the theorem.

REMARK 4.1. To show a similar impossibility result for general dense regular graphs, we need to get exact upper bounds on $Z_n(\beta,B)$. More precisely, we need to show that $Z_n(\beta,B) \le O(1) Z_n^{\mathrm{cw}}(\beta,B)$.

4.2. *Proof of Theorem* 1.8. For proving Theorem 1.8, we need the following two lemmas. The proofs of the lemmas are deferred to the end of the section. The first lemma gives an estimate similar to Lemma 2.1.

LEMMA 4.2. *Suppose* $\mathbf{X} = (X_1,\dots,X_n)$ *is an observation from the Ising model* (1.1), *where the coupling matrix* A_n *satisfies* (1.2) *and* (1.3), *and* $\beta > 0$, $B \neq 0$. *Then we have*

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n \left(m_i(\mathbf{X}) - \sum_{j=1}^n A_n(i,j)\tanh\bigl(\beta m_j(\mathbf{X}) + B\bigr)\right)\right]^2 < \infty.$$

The second lemma proves an estimate for the Curie–Weiss model, which is necessary for completing the proof of Theorem 1.8.

LEMMA 4.3. *Let* β_n *be a sequence of positive reals bounded away from* 0 *and* $+\infty$, *and* A_n *be a matrix satisfying* (1.2) *and* (1.9). *Then for any* $B \in \mathbb{R}$ *there exists* $\delta > 0$ *such that under the Curie–Weiss model* $\mathbb{P}^{\mathrm{cw}}_{n,\beta_n,B}$ *we have*

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}^{\mathrm{cw}}_{n,\beta_n,B}\left(\sum_{i=1}^n \bigl(m_i(\mathbf{X}) - \bar{m}(\mathbf{X})\bigr)^2 \le n\delta\right) < 0,$$

where $m(\mathbf{X}) := A_n\mathbf{X}$.

PROOF OF THEOREM 1.8. It suffices to show that given an arbitrary sequence of positive reals $\{\varepsilon_n\}_{n\ge1}$ we have

$$\lim_{n\to\infty} \mathbb{P}_{n,\beta,B}\left(\sum_{i=1}^n \bigl(m_i(\mathbf{X}) - \bar{m}(\mathbf{X})\bigr)^2 \le n\varepsilon_n\right) = 0.$$

Assume by way of contradiction that

$$\limsup_{n\to\infty} \mathbb{P}_{n,\beta,B}(E_n) > 0, \qquad E_n := \left\{\sum_{i=1}^n \bigl(m_i(\mathbf{X}) - \bar{m}(\mathbf{X})\bigr)^2 \le n\epsilon_n\right\}.$$

Using (2.3) and Lemma 4.2 and increasing the value of ε_n if necessary, we have $\lim_{n\to\infty}\mathbb{P}_{n,\beta,B}(C_n\cap D_n)=1$, where

$$C_n:=\left\{\sum_{i=1}^n\bigl(X_i-\tanh\bigl(\beta m_i(\mathbf{X})+B\bigr)\bigr)\le n\varepsilon_n\right\},$$

$$D_n:=\left\{\sum_{i=1}^n\left(m_i(\mathbf{X})-\sum_{j=1}^n A_n(i,j)\tanh\bigl(\beta m_j(\mathbf{X})+B\bigr)\right)\le n\varepsilon_n\right\}.$$

Combining this gives $\limsup_{n\to\infty}\mathbb{P}_{n,\beta,B}(C_n\cap D_n\cap E_n)>0$, and so there is a subsequence along which $\mathbb{P}_{n,\beta,B}(C_n\cap D_n\cap E_n)\ge\delta$ for some $\delta>0$. Restricting ourself to this subsequence, on the set $C_n\cap D_n\cap E_n$ we have

$$\begin{aligned}\sum_{i=1}^n X_i &\overset{C_n}{=}\sum_{i=1}^n\tanh\bigl(\beta m_i(\mathbf{X})+B\bigr)+o(n)\\ &\overset{E_n}{=} n\tanh\bigl(\beta\bar{m}(\mathbf{X})+B\bigr)+o(n),\\ n\bar{\mathcal{R}}_n\tanh\bigl(\beta\bar{m}(\mathbf{X})+B\bigr) &\overset{E_n}{=}\sum_{i=1}^n\sum_{j=1}^n A_n(i,j)\tanh\bigl(\beta m_i(\mathbf{X})+B\bigr)+o(n)\\ &\overset{D_n}{=}\sum_{i=1}^n m_i(\mathbf{X})+o(n).\end{aligned}$$

In the above sequence of equations, the $o(n)$ terms are uniform nonrandom bounds over the set $C_n\cap D_n\cap E_n$ which on dividing by n go to 0 as $n\to\infty$, and are not made explicit for the sake of clarity. On combining the above two equations, we get $\bar{\mathcal{R}}_n\sum_{i=1}^n X_i=\sum_{i=1}^n m_i+o(n)$, using which gives

$$\sum_{i=1}^n X_i m_i(\mathbf{X})\overset{E_n}{=}\bar{m}\sum_{i=1}^n x_i=n\bar{\mathcal{R}}_n\bar{X}^2+o(n).$$

Thus, we have

$$\begin{aligned}\delta&\le\mathbb{P}_{n,\beta,B}(C_n\cap D_n\cap E_n)\\ &=\frac{1}{Z_n(\beta,B)}\sum_{\mathbf{x}\in C_n\cap D_n\cap E_n}e^{\frac{\beta}{2}\mathbf{x}'A_n\mathbf{x}+B\sum_{i=1}^n x_i}\\ &\le\frac{e^{o(n)}}{Z_n(\beta,B)}\sum_{\mathbf{x}\in C_n\cap D_n\cap E_n}e^{\frac{\beta\bar{\mathcal{R}}_n}{2}\bar{x}^2+nB\bar{x}}\\ &=e^{o(n)}\frac{Z_n^{\mathrm{cw}}(\beta_n,B)}{Z_n(\beta,B)}\mathbb{P}^{\mathrm{cw}}_{n,\beta_n,B}(C_n\cap D_n\cap E_n),\end{aligned}$$

where $\beta_n:=\beta\bar{\mathcal{R}}_n$ is a sequence of positive real bounded away from ∞ and 0 by (1.2) and (1.9) respectively, and $\mathbb{P}^{\mathrm{cw}}_{n,\beta_n,B}$ is the Curie–Weiss model with parameters (β_n,B). Now, using [2], (1.8), and [2], (2.3), we get

$$\log Z_n(\beta,B)\ge n\sup_{t\in[-1,1]}\left\{\frac{\beta}{2}\bar{\mathcal{R}}_n t^2+Bt-I(t)\right\},$$

$$\log Z_n^{\mathrm{cw}}(\beta_n,B)=n\sup_{t\in[-1,1]}\left\{\frac{\beta}{2}\bar{\mathcal{R}}_n t^2+Bt-I(t)\right\}+o(n)$$

respectively, which readily gives

$$\delta \geq e^{o(n)}\mathbb{P}^{\text{cw}}_{n,\beta_n,B}(C_n \cap D_n \cap E_n) \leq e^{o(n)}\mathbb{P}^{\text{cw}}_{n,\beta_n,B}(E_n),$$

which on taking log, dividing by n and letting $n \to \infty$ gives

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}^{\text{cw}}_{n,\beta_n,B}(E_n) = 0.$$

But this is a contradiction to Lemma 4.3, which completes the proof of the theorem. □

PROOF OF LEMMA 4.2. To begin, note that

$$\sum_{i=1}^n \biggl(m_i(\mathbf{X}) - \sum_{j=1}^n A_n(i,j)\tanh\bigl(\beta m_j(\mathbf{X}) + B\bigr)\biggr) = \sum_{j=1}^n \mathcal{R}_n(j)((X_j - b_j(\mathbf{X})),$$

where $b_j(\mathbf{x}) = \tanh(\beta m_i(\mathbf{x}) + B)$ as in Lemma 2.1. This on squaring and expanding gives

$$\begin{aligned} &\mathbb{E}\Biggl[\sum_{i=1}^n \biggl(m_i(\mathbf{X}) - \sum_{j=1}^n A_n(i,j)\tanh\bigl(\beta m_j(\mathbf{X}) + B\bigr)\biggr)\Biggr]^2 \\ &\quad = \sum_{j=1}^n \mathcal{R}_n(j)^2 \mathbb{E}\bigl(X_j - b_j(\mathbf{X})\bigr)^2 \\ &\qquad + \sum_{j\neq k} \mathcal{R}_n(j)\mathcal{R}_n(k)\mathbb{E}\bigl(X_j - b_j(\mathbf{X})\bigr)\bigl(X_k - b_k(\mathbf{X})\bigr), \end{aligned} \tag{4.1}$$

where the first term in (4.1) is bounded by $n\gamma^2$ by (1.2). Proceeding to bound the second term, setting $m_j^{(i)}(\mathbf{X}) = \sum_{k\neq i} A(j,k)x_k$ we have

$$\begin{aligned} &\mathbb{E}\bigl(X_i - \tanh\bigl(\beta m_i(\mathbf{X}) + B\bigr)\bigr)\bigl(X_j - \tanh\bigl(\beta m_j^{(i)}(\mathbf{X}) + B\bigr)\bigr) \\ &\quad = \mathbb{E}\bigl\{\mathbb{E}\bigl(X_i - \tanh\bigl(\beta m_i(\mathbf{X}) + B\bigr)\bigr)|\{X_k, k \neq i\}\bigr)\bigl(X_j - \tanh\bigl(\beta m_j^{(i)}(\mathbf{X}) + B\bigr)\bigr)\bigr\} \\ &\quad = 0, \end{aligned}$$

which gives

$$\begin{aligned} &\bigl|\mathbb{E}\bigl(X_i - \tanh\bigl(\beta m_i(\mathbf{X}) + B\bigr)\bigr)\bigl(X_j - \tanh\bigl(\beta m_j(\mathbf{X}) + B\bigr)\bigr)\bigr| \\ &\quad = \bigl|\mathbb{E}\bigl(X_i - \tanh\bigl(\beta m_i(\mathbf{X}) + B\bigr)\bigr)\bigl(\tanh\bigl(\beta m_j^{(i)}(\mathbf{X}) + B\bigr) - \tanh\bigl(\beta m_j(\mathbf{X}) + B\bigr)\bigr)\bigr| \\ &\quad \leq \beta A_n(i,j). \end{aligned}$$

Summing over $i \neq j$, the second term in (4.1) is bounded by

$$\sum_{i\neq j} \mathcal{R}_n(i)\mathcal{R}_n(j)\beta A_n(i,j) \leq n\beta\gamma^3.$$

Using (4.1) and combining we get

$$\mathbb{E}\Biggl[\sum_{i=1}^n \biggl(m_i(\mathbf{X}) - \sum_{j=1}^n A_n(i,j)\tanh\bigl(\beta m_j(\mathbf{X}) + B\bigr)\biggr)\Biggr]^2 \leq n\gamma^2 + n\beta\gamma^3. \tag{4.2}$$

This completes the proof of the lemma. □

PROOF OF LEMMA 4.3. To begin, use [19], Lemma 3, to note that there exists a random variable W_n, such that given $W_n = w$ we have $(X_1, \dots, X_n)$ are i.i.d. with

$$\mathbb{P}^{\text{cw}}_{n,\beta_n,B}(X_i = 1) = \frac{e^{\beta_n w + B}}{e^{\beta_n w + B} + e^{-\beta_n w - B}} = 1 - \mathbb{P}^{\text{cw}}_{n,\beta_n,B}(X_i = -1).$$

Also note that

$$\begin{aligned} Y_n(\mathbf{x}) &:= \sqrt{\sum_{i=1}^n \bigl(m_i(\mathbf{x}) - \bar{m}(\mathbf{x})\bigr)^2} = \sup_{\|\mathbf{a}\|_2 \le 1} a_i\bigl(m_i(\mathbf{x}) - \bar{m}(\mathbf{x})\bigr) \\ &= \sup_{\|\mathbf{a}\|_2 \le 1} \mathbf{a}'\Bigl(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\Bigr)A_n\mathbf{x} \end{aligned}$$

and so $Y_n(\mathbf{x})$ is a convex function of $\mathbf{x}$. Furthermore, for any $\mathbf{y}$ and $\mathbf{z}$, using the Cauchy–Schwarz inequality we get $|Y_n(\mathbf{y}) - Y_n(\mathbf{z})| \le \gamma\|\mathbf{y} - \mathbf{z}\|$, where we use the bound $\|A_n\|_2 \le \gamma$. Therefore, $Y_n(\mathbf{x})$ is a γ-Lipschitz function of $\mathbf{x}$ on $[-1, 1]^n$, and so invoking [6], Theorem 7.12, gives

$$\mathbb{P}\bigl(Y_n(\mathbf{X}) \le \mathbb{E}\bigl(Y_n(\mathbf{X})|W_n\bigr) - 2\gamma\sqrt{t}|W_n\bigr) \le 2e^{-\frac{t}{4}}. \tag{4.3}$$

In order to invoke (4.3), we need to first estimate $\mathbb{E}(Y_n|W_n)$. To this effect, a direct computation gives

$$\begin{aligned} &\mathbb{E}\bigl(Y_n(\mathbf{X})|W_n\bigr) \\ &= \mathbb{E}\Biggl(\sqrt{\Biggl(\sum_{i=1}^n m_i(\mathbf{X}) - \bar{m}(\mathbf{X})\Biggr)^2}\Bigg|W_n\Biggr) \\ &\ge \frac{1}{\sqrt{n}}\mathbb{E}\Biggl(\sum_{i=1}^n \bigl|m_i(\mathbf{X}) - \bar{m}(\mathbf{X})\bigr|\Bigg|W_n\Biggr) \quad \text{[By Cauchy–Schwarz]} \\ &\ge \frac{1}{2\gamma\sqrt{n}}\mathbb{E}\Biggl(\sum_{i=1}^n \bigl(m_i(\mathbf{X}) - \bar{m}(\mathbf{X})\bigr)^2\Bigg|W_n\Biggr) \quad \bigl[\text{Since } \bigl|m_i(\mathbf{x}) - \bar{m}(\mathbf{x})\bigr| \le 2\gamma\bigr] \\ &\ge \frac{1}{2\gamma\sqrt{n}}\sum_{i=1}^n \operatorname{Var}\bigl(m_i(\mathbf{X}) - \bar{m}(\mathbf{X})|W_n\bigr) \\ &= \frac{1}{2\gamma\sqrt{n}}\operatorname{Var}(X_1|W_n)\sum_{i,j=1}^n \Bigl(A_n(i,j) - \frac{1}{n}\mathcal{R}_n(i)\Bigr)^2. \end{aligned}$$

Finally, we have

$$\sum_{i,j=1}^n \Bigl(A_n(i,j) - \frac{1}{n}\mathcal{R}_n(i)\Bigr)^2 = \sum_{i,j=1}^n A_n^2(i,j) - \frac{1}{n}\sum_{i=1}^n \mathcal{R}_n(i)^2 \ge \sum_{i,j=1}^n A_n^2(i,j) - \gamma^2,$$

where the last inequality follows by (1.2). Since $\operatorname{Var}(X_1|W_n) = \operatorname{sech}^2(\beta_n W_n + B)$, on the set $|W_n| \le 2$ we have

$$\mathbb{E}(Y_n|W_n) \ge \frac{1}{2\gamma\sqrt{n}}\operatorname{sech}^2(\beta_n W_n + B)\Biggl(\sum_{i,j=1}^n A_n^2(i,j) - \gamma\Biggr) \ge 2c\sqrt{n}$$

for some constant $c > 0$ by invoking (1.9). Thus on the set $|W_n| \le 2$ invoking (4.3) with $t = \frac{nc^2}{4\gamma^2}$ gives

$$\mathbb{P}\big(Y_n(\mathbf{X}) \le c\sqrt{n}|W_n\big) \le \mathbb{P}\big(Y_n(\mathbf{X}) \le 2c\sqrt{n} - 2\gamma\sqrt{t}|W_n\big) \le 2e^{-\frac{t}{4}} = 2e^{-\frac{nc^2}{16\gamma^2}}.$$

With $\delta = c^2$, this gives

$$\begin{aligned}&\mathbb{P}^{\mathrm{cw}}_{n,\beta_n,B}\left(\sum_{i=1}^{n}\big(m_i(\mathbf{X}) - \bar{m}(\mathbf{X})\big)^2 \le n\delta\right)\\ &\quad\le \mathbb{P}^{\mathrm{cw}}_{n,\beta_n,B}\big(|W_n| > 2\big) + \mathbb{E}^{\mathrm{cw}}_{n,\beta_n,B}\big(\mathbb{P}^{\mathrm{cw}}_{n,\beta_n,B}\big(Y_n(\mathbf{X}) \le c\sqrt{n}|W_n\big)1\{|W_n| \le 2\}\big)\\ &\quad\le \mathbb{P}^{\mathrm{cw}}_{n,\beta_n,B}\big(|W_n| > 2\big) + 2e^{-\frac{nc^2}{16\gamma^2}},\end{aligned}$$

from which the desired conclusion follows on using [19], Lemma 3, to note that the first term in the RHS above decays exponentially, as $(W_n|X_1, \dots, X_n) \sim N(\bar{X}, \frac{1}{n\beta_n})$. □

Acknowledgments. We thankfully acknowledge helpful discussions with Bhaswar B. Bhattacharya and Sourav Chatterjee at various stages of this work. We would also like to thank the Associate Editor and two anonymous referees, whose input and suggestions have greatly benefited the presentation of this paper.

This work was supported in part by the NSF Grants DMS-1664650 and DMS-1712037.

SUPPLEMENTARY MATERIAL

Supplement A: Appendix (DOI: 10.1214/19-AOS1822SUPP; .pdf). Supplement A contains the proofs of Theorem 3.2, Lemma 2.2 and Proposition 1.9.

REFERENCES

[1] ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). High-dimensional structure estimation in Ising models: Local separation criterion. *Ann. Statist.* **40** 1346–1375. MR3015028 https://doi.org/10.1214/12-AOS1009

[2] BASAK, A. and MUKHERJEE, S. (2017). Universality of the mean-field for the Potts model. *Probab. Theory Related Fields* **168** 557–600. MR3663625 https://doi.org/10.1007/s00440-016-0718-0

[3] BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

[4] BESAG, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc., Ser. D Stat.* **24** 179–195.

[5] BHATTACHARYA, B. B. and MUKHERJEE, S. (2018). Inference in Ising models. *Bernoulli* **24** 493–525. MR3706767 https://doi.org/10.3150/16-BEJ886

[6] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[7] BRESLER, G. (2015). Efficiently learning Ising models on arbitrary graphs. In *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing* 771–782. ACM, New York. MR3388257

[8] CHATTERJEE, S. (2007). Estimation in spin glasses: A first step. *Ann. Statist.* **35** 1931–1946. MR2363958 https://doi.org/10.1214/009053607000000109

[9] CHATTERJEE, S. and DEMBO, A. (2016). Nonlinear large deviations. *Adv. Math.* **299** 396–450. MR3519474 https://doi.org/10.1016/j.aim.2016.05.017

[10] COMETS, F. (1992). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *Ann. Statist.* **20** 455–468. MR1150354 https://doi.org/10.1214/aos/1176348532

[11] COMETS, F. and GIDAS, B. (1991). Asymptotics of maximum likelihood estimators for the Curie–Weiss model. *Ann. Statist.* **19** 557–578. MR1105836 https://doi.org/10.1214/aos/1176348111

[12] DEMBO, A. and MONTANARI, A. (2010). Gibbs measures and phase transitions on sparse random graphs. *Braz. J. Probab. Stat.* **24** 137–211. MR2643563 https://doi.org/10.1214/09-BJPS027

[13] ELDAN, R. and GROSS, R. (2018). Decomposition of mean-field Gibbs distributions into product measures. *Electron. J. Probab.* **23** Paper No. 35. MR3798245 https://doi.org/10.1214/18-EJP159

[14] ELLIS, R. S. and NEWMAN, C. M. (1978). The statistics of Curie–Weiss models. *J. Stat. Phys.* **19** 149–161. MR0503332 https://doi.org/10.1007/BF01012508

[15] GHOSAL, P. and MUKHERJEE, S. (2020). Supplement to "Joint estimation of parameters in Ising model." https://doi.org/10.1214/19-AOS1822SUPP.

[16] GIDAS, B. (1988). Consistency of maximum likelihood and pseudolikelihood estimators for Gibbs distributions. In *Stochastic Differential Systems*, *Stochastic Control Theory and Applications* (*Minneapolis*, *Minn*., 1986). *IMA Vol. Math. Appl.* **10** 129–145. Springer, New York. MR0934721 https://doi.org/10.1007/978-1-4613-8762-6_10

[17] GUYON, X. and KÜNSCH, H. R. (1992). Asymptotic comparison of estimators in the Ising model. In *Stochastic Models*, *Statistical Methods*, *and Algorithms in Image Analysis* (*Rome*, 1990). *Lect. Notes Stat.* **74** 177–198. Springer, Berlin. MR1188486 https://doi.org/10.1007/978-1-4612-2920-9_12

[18] LOVÁSZ, L. (2012). *Large Networks and Graph Limits. American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035 https://doi.org/10.1090/coll/060

[19] MUKHERJEE, R., MUKHERJEE, S. and YUAN, M. (2018). Global testing against sparse alternatives under Ising models. *Ann. Statist.* **46** 2062–2093. MR3845011 https://doi.org/10.1214/17-AOS1612

[20] RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 https://doi.org/10.1214/09-AOS691

[21] XUE, L., ZOU, H. and CAI, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *Ann. Statist.* **40** 1403–1429. MR3015030 https://doi.org/10.1214/12-AOS1017