STAT 9700 Project: Estimation in Ising models

Sunay Joshi

December 11, 2024

Abstract

In this paper, we discuss estimation in Ising models. We begin by presenting the results of Chatterjee (2007) in the case of zero external field, using Stein's method of exchangeable pairs to prove $\sqrt{n}$-consistency. Then we present the generalization to nonzero external fields under low-rank structure given by Daskalakis et al. (2019). We also present an impossibility result in the setting of constant external field given by Ghosal and Mukherjee (2020), which relies on contiguity arguments. Finally, we discuss the follow-up work Kandiros et al. (2021).

1 Introduction

In statistical physics, the Ising model is one of the simplest models of ferromagnetism. The magnet is modeled as a collection of n sites, each with its own magnetic spin σ_i that can either point up ($\sigma_i = +1$) or down ($\sigma_i = -1$). The collection of all spins is called a configuration $\sigma \in \{\pm 1\}^n$. The model is a probability distribution over configurations σ specified by an $n \times n$ symmetric coupling matrix A_n, an inverse temperature $\beta > 0$, and external magnetic fields h_i at each site, and is given by the exponential family

$$\mathbb{P}_{n,\beta,h}(\sigma) = \frac{1}{Z_n} \exp\left(\frac{1}{2}\beta\sigma^T A_n \sigma + \sum_i \sigma_i h_i\right),$$

where $Z_n = Z_n(\beta, h)$ denotes the partition function. In the limit as $\beta \to 0$, the spins σ_i become independent, but at finite temperature, the spins can have significant correlations due to the entries of the coupling matrix A. Physically, this gives rise to a ferromagnetic phase transition at a certain critical temperature. In this paper, we discuss the difficulties that such dependencies pose to the estimation of the parameters β and h_i from a single sample $\sigma \sim \mathbb{P}_{n,\beta,h}$. The first obstacle is that the maximum likelihood estimator is computationally intractable, as the log-partition function (a sum over 2^n configurations) is difficult to compute. As a result, we instead turn to the maximum pseudolikelihood estimator (MPLE) Besag (1974).

In this paper, we consider estimation in a number of scenarios. We begin with the simplest setting, when $h_i = 0$ for all $i \in [n]$, in which case we only must estimate β. Here the results of Chatterjee (2007) show that estimation is possible at the $\sqrt{n}$-rate with the MPLE under certain

assumptions on the coupling matrix A_n. The proof utilizes Stein's method of exchangeable pairs, a tool to prove concentration inequalities in dependent settings. Next, we discuss more general structures of external fields (h_i). We sketch the results of Daskalakis et al. (2019), which show that when $h_i = x_i^T\theta$ for known features $x_i \in \mathbb{R}^d$ and unknown $\theta \in \mathbb{R}^d$, then the parameters (β, θ) can again be estimated at the $\sqrt{n}$-rate under similar conditions to Chatterjee (2007). This can be thought of as a low-rank assumption on the external fields, or as an example of networked logistic regression with covariates (x_i). We then present a result of a different flavor from Kandiros et al. (2021), which proves that in an Erdos-Renyi Ising model with constant external field $h_i = B$, consistent estimation of (β, B) is impossible in a certain region of the parameter space. The proof utilizes the notion of contiguity of probability measures, a concept that is useful to prove inconsistency or nonexistence of powerful tests. Finally, we discuss the follow-up paper Kandiros et al. (2021) that relaxes the assumptions of Daskalakis et al. (2019).

Regarding notation: all sums run from 1 to n, we at times drop reference to the number of spins n, and we refer to the true parameters as (β, θ).

2 Main results

2.1 Zero external field

We begin with Chatterjee (2007) in the case of zero external field.

Let $\mathbb{P}_{n,\beta}$ denote the Ising model at inverse temperature $\beta > 0$ with zero external field. We seek to estimate the unknown parameter β from a single sample σ drawn from $\mathbb{P}_{n,\beta}$. The MPLE will be denoted $\hat{\beta}_n$ below. The pseudolikelihood (PL) is given by the product

$$\mathrm{PL}(\beta) = \prod_i \mathbb{P}_{n,\beta}(\sigma_i | \sigma_j, j \neq i) = \prod_i \frac{\exp(\sigma_i m_i(\sigma))}{\exp(\sigma_i m_i(\sigma)) + \exp(-\sigma_i m_i(\sigma))},$$

and the MPLE is simply defined as its argmax:

$$\hat{\beta}_n = \mathrm{argmax}_\beta \, \mathrm{PL}(\beta).$$

Here we define the local field at site i as

$$m_i(\sigma) = \sum_j A_n(i,j)\sigma_j,$$

which is the effective external magnetic field the site i, thought of as a single-spin system, experiences from its surroundings. The PL score is the derivative of the log-PL with respect to β and is given by

$$\nabla \log \mathrm{PL}(\beta) = \frac{1}{n}\sum_i m_i(\sigma)(\sigma_i - \tanh(\beta m_i(\sigma))).$$

The MPLE is a zero of the PL score. The first result we discuss is (Chatterjee, 2007, Theorem 1.1), which shows that under certain conditions on the coupling matrix and on β, the MPLE is $\sqrt{n}$-consistent. Let $\|M\|_2$ denote the induced 2-norm of a matrix M.

Theorem 2.1. *Assume that for a given $\beta > 0$, we have $\sup_n \|A_n\|_2 < \infty$ and $\liminf_n \frac{1}{n} \log Z_n > \log 2$. Then the MPLE $\hat{\beta}_n$ is $\sqrt{n}$-consistent for β.*

The proof of Theorem 2.1 involves two steps. First, it is shown that the true β is a near-zero of the PL score $\nabla \log \mathrm{PL}$. Then, it is shown with high probability that $\nabla \log \mathrm{PL}$ is strongly concave in a certain region, from which it follows that β must be close to the global minimizer $\hat{\beta}_n$. The first step is the following second moment bound from (Chatterjee, 2007, Lemma 1.1).

Lemma 2.2. *Given $\beta \geqslant 0$, let ∇PL denote the PL score with $\sigma \sim \mathbb{P}_{n,\beta}$. Then*

$$\mathbb{E}[\nabla \log PL(\beta)^2] \leqslant \frac{C}{n},$$

where $C = 6\|A_n\|_2^2 + 6\beta\|A_n\|_2^3 + 2\beta^2\|A_n\|_2^4$.

Proof. By rescaling, we may assume that $\|A_n\|_2 = 1$. To bound the second moment of $\nabla \log \mathrm{PL}(\beta)$, we rely on a stability argument, relating the variability of $\nabla \log \mathrm{PL}(\beta)$ to its sensitivity to perturbations in σ. This technique is known as Stein's method of exchangeable pairs; see Chatterjee (2005). It is reminiscent of the "replace one" procedure used to prove concentration inequalities like the Efron-Stein inequality, which states that for a function $f : \mathbb{R}^n \to \mathbb{R}$ and a random vector $X = (X_1, \ldots, X_n)$ with independent components, we may bound its variance by the decomposition

$$\mathrm{Var} f(X) \leqslant \frac{1}{2} \sum_i \mathbb{E}[(f(X^{(i)}) - f(X))^2],$$

where $X^{(i)}$ denotes the random vector obtained by independently resampling the i-th component of X Chatterjee (2005). By showing that each of the terms on the right-hand side are small, we deduce that $f(X)$ concentrates around its mean.

In our case, σ has highly dependent entries. Our notion of perturbation comes as a single step of Glauber dynamics. That is, the perturbation σ' of σ is obtained by sampling a site uniformly at random and resampling its spin from the conditional distribution of σ_i given σ_j for $j \neq i$.

To utilize this notion of perturbation, we will write our quantity of interest $\nabla \log \mathrm{PL}(\beta)^2$ in terms of $f(\sigma) - f(\sigma')$ for some function f. Since

$$\mathbb{E}[\sigma_i | \sigma_j, j \neq i] = \frac{\exp(\sigma_i m_i(\sigma))}{\exp(\sigma_i m_i(\sigma)) + \exp(-\sigma_i m_i(\sigma))} - \frac{\exp(\sigma_i m_i(-\sigma))}{\exp(\sigma_i m_i(\sigma)) + \exp(-\sigma_i m_i(\sigma))} = \tanh(\beta m_i(\sigma)),$$

we may write

$$\nabla \log \mathrm{PL}(\beta) = \frac{1}{n} \sum_i m_i(\sigma)(\sigma_i - \mathbb{E}[\sigma_i | \sigma_j, j \neq i]).$$

Replacing the average with an expectation over a uniformly random index I, we obtain

$$\nabla \log \mathrm{PL}(\beta) = \mathbb{E}[m_I(\sigma)(\sigma_I - \mathbb{E}[\sigma_I | \sigma_j, j \neq I]) | \sigma],$$

and letting σ' denote the perturbation of σ as defined above, we obtain

$$\nabla \log \mathrm{PL}(\beta) = \mathbb{E}[m_I(\sigma)(\sigma_I - \sigma'_I)|\sigma].$$

We have arrived at $\nabla \log \mathrm{PL}(\beta) = f(\sigma) = \mathbb{E}\left[F(\sigma, \sigma')|\sigma\right]$, where the antisymmetric function F is defined as

$$F(\tau, \tau') = \frac{1}{2}\sum_i (m_i(\tau) + m_i(\tau'))(\tau - \tau'_i).$$

We may now relate the second moment of $f(\sigma)$ to the stability of $f(\sigma)$ via the manipulation

$$\begin{aligned}
\mathbb{E}\left[f(\sigma)^2\right] &= \mathbb{E}\left[f(\sigma)\mathbb{E}\left[F(\sigma, \sigma')|\sigma\right]\right] \\
&= \mathbb{E}\left[f(\sigma)F(\sigma, \sigma')\right] \\
&= \mathbb{E}\left[f(\sigma')F(\sigma', \sigma)\right] \\
&= \frac{1}{2}\mathbb{E}\left[(f(\sigma) - f(\sigma'))F(\sigma, \sigma')\right],
\end{aligned}$$

where in the last step we used exchangeability and antisymmetry.

Consider the conditional expectation of this quantity given σ. Each index is equally likely to be resampled, and the random variable is nonzero if and only if the sign of the chosen spin is flipped. Using our formula for $F(\sigma, \sigma')$ we deduce

$$\mathbb{E}\left[(f(\sigma) - f(\sigma'))F(\sigma, \sigma')|\sigma\right] = \frac{1}{n}\sum_j (f(\sigma) - f(\sigma^{(j)}))2m_j(\sigma)\sigma_j p_j(\sigma),$$

where $p_j(\sigma)$ is the conditional probability that $\sigma_j^{(j)} = -1$ given the remaining spins.

We must understand the quantity $f(\sigma) - f(\sigma^{(j)})$, where $\sigma^{(j)}$ denotes the configuration obtained by flipping site j. From the definition of f, we obtain

$$\begin{aligned}
f(\sigma) - f(\sigma^{(j)}) &= \frac{1}{n}\sum_i (m_i(\sigma) - m_i(\sigma^{(j)}))(\sigma_i - \tanh(\beta m_i(\sigma))) + \frac{1}{n}\sum_i m_i(\sigma^{(j)})(\sigma_i - \sigma_i^{(j)}) \\
&\quad - \frac{1}{n}\sum_i m_i(\sigma^{(j)})(\tanh(\beta m_i(\sigma)) - \tanh(\beta m_i(\sigma^{(j)}))) \\
&= \frac{2\sigma_j}{n}\sum_i A_{ij}(\sigma_i - \tanh(\beta m_i(\sigma))) + \frac{2}{n}m_j(\sigma)\sigma_j \\
&\quad - \frac{1}{n}\sum_i m_i(\sigma^{(j)})(\tanh(\beta m_i(\sigma)) - \tanh(\beta m_i(\sigma^{(j)}))) \\
&=: (I) + (II) + (III)
\end{aligned}$$

Finally, we plug these back into our formula for the conditional second moment. As it turns out, all three terms are on the order of $1/n$, but we simply show this for Term (II). Using $\sigma_j^2 = 1$, $p_j(\sigma) \leqslant 1$, and $|m_j(\sigma)| \leqslant \|A_n e_j\|_2 \leqslant \|A_n\|_2 = 1$, Term (II) contributes

$$\frac{1}{n^2}\sum_j 4m_j(\sigma)^2 p_j(\sigma) \leqslant \frac{1}{n^2} \cdot 4n = \frac{4}{n}.$$

Putting everything together yields the second moment bound. □

Now we must show that $\nabla \log \mathrm{PL}(\beta) \approx 0$ implies β is close to the zero $\hat{\beta}_n$ of $\nabla \log \mathrm{PL}$. As discussed above, this relies on strong concavity of the log-PL function, and is shown in (Chatterjee, 2007, Lemmas 2.2-2.6). We simply sketch (Chatterjee, 2007, Lemma 2.2), rephrased below.

Lemma 2.3. *Define the Hamiltonian of a configuration as the quadratic form $H_n(\tau) = \frac{1}{2}\tau^T A_n \tau$. If $n^{-1}H_n(\tau) \geqslant c > 0$ for a given configuration τ, then the log-PL is concave with*

$$\nabla^2 \log PL(x) \leqslant -\frac{3}{4}c^4 sech^2(2x/c),$$

from which it follows that

$$|\tanh(2\hat{\beta}_n(\tau)/c) - \tanh(2\beta/c)| \leqslant \frac{8|\nabla \log PL(\beta)|}{3c^5}.$$

Proof. (Sketch.) We begin by writing out the derivative of the score:

$$\nabla^2 \log \mathrm{PL}(x) = -\frac{1}{n}\sum_i m_i(\tau)^2 \mathrm{sech}^2(x m_i(\tau)).$$

It is immediate from this expression the log-PL is concave. The lower bound on the Hamiltonian allows us to make this quantitative. Since $H_n(\tau) = \sum_i m_i(\tau)\tau_i$, the assumption $n^{-1}H_n(\tau) \geqslant c$ implies that the local fields $m_i(\tau)$'s can't all be small in absolute value. Specifically, one can show that for at least a fraction $3c^2/4$ of the sites, the magnitude of the local field $|m_i(\sigma)|$ lies in the interval $[c, 2/c]$. Plugging this back into the expression for $\nabla^2 \log \mathrm{PL}(x)$ and optimizing, we obtain the first inequality. Integrating $\nabla^2 \log \mathrm{PL}(x)$ from $x = \hat{\beta}_n$ to $x = \beta$, we obtain a lower bound on $|\nabla \log \mathrm{PL}(\beta)| = |\nabla \log \mathrm{PL}(\beta) - \nabla \log \mathrm{PL}(\hat{\beta}_n)|$, and rearranging yields the second inequality. □

2.2 Low-rank external field

Next, we discuss Daskalakis et al. (2019), a natural generalization of Chatterjee (2007) is to consider Ising models with nonzero external fields. Here we have features $x_i \in \mathbb{R}^d$ associated to each site $i \in [n]$, the external field at site i is given by $h_i = x_i^T\theta$, and we seek to estimate the unknown parameters $\beta \in \mathbb{R}$ and $\theta \in \mathbb{R}^d$. Given a sample $\sigma \sim \mathbb{P}_{n,\beta,\theta}$ from this model, the MPLE $(\hat{\theta}_n, \hat{\beta}_n)$ can be defined as above. The following result, (Daskalakis et al., 2019, Theorem 3.1), shows that under similar conditions to Chatterjee (2007), the MPLE is $\sqrt{n}$-consistent for both θ and β.

Theorem 2.4. *Suppose that $\beta \in (-b, b)$, $\theta \in [-\Theta, \Theta]^d$, and $x_i \in [-M, M]^d$ for all $i \in [n]$. Suppose $\|A_n\|_\infty \leqslant 2$ and $\|A_n\|_F^2 \geqslant cn$. Let X denote the design matrix whose i-th row equals x_i^T, let $Q = \frac{1}{n}X^T X$, suppose Q is positive definite, and suppose $\lambda_{min}(Q), \lambda_{max}(Q) = \Theta(1)$. Then for each $\delta > 0$, for sufficiently large n, the MPLE $(\hat{\theta}_n, \hat{\beta}_n)$ satisfies*

$$\|(\hat{\theta}_n, \hat{\beta}_n) - (\theta, \beta)\|_2 \leqslant O\left(\frac{1}{\sqrt{n}}\right)$$

with probability at least $1 - \delta$.

Proof. (Sketch.) The proof follows quite a similar outline to Chatterjee (2007), except the authors consider a multidimensional parameter. First, the authors show that the true parameters (θ, β) are a near-zero of the PL-score. Then, the authors show that the negative log-PL is strongly convex, from which it follows that (θ, β) is close to $(\hat{\theta}_n, \hat{\beta}_n)$.

We use the following fact from convex analysis.

Lemma 2.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a λ-strongly convex function. Let x^* denote its global minimum. Then*

$$\lambda \|x - x^*\|_2 \leqslant \|\nabla f(x)\|_2.$$

From a direct computation, the Hessian of the negative log-PL is bounded between

$$\frac{1}{n} \sum_i X_i X_i^T \succeq -\nabla^2 \log \mathrm{PL} \succeq \operatorname{sech}^2(B + dM\Theta) \cdot \frac{1}{n} \sum_i X_i X_i^T,$$

and in particular is convex. Further, by definition, the negative log-PL is $\lambda_{\min}(-\nabla^2 \log \mathrm{PL})$-convex. Thus by Theorem 2.5, if $(\hat{\theta}_n, \hat{\beta}_n) \in [-\Theta, \Theta]^d \times (-B, B)$, then we have

$$\lambda_{\min}(-\nabla^2 \log \mathrm{PL}) \cdot \|(\hat{\theta}_n, \hat{\beta}_n) - (\theta, \beta)\|_2 \leqslant \|\nabla \log \mathrm{PL}(\theta, \beta)\|_2.$$

By (Daskalakis et al., 2019, Lemma 3.2-3.3), the gradient of the score at the true parameters (θ, β) has small norm with high probability; in particular, $\|\nabla \log \mathrm{PL}(\theta, \beta)\| = O(1/\sqrt{n})$. (This is the step that requires Stein's method of exchangeable pairs.) Next, (Daskalakis et al., 2019, Lemma 3.4) shows that under the assumption $\|A_n\|_F^2 \geqslant cn$, the negative log-PL is λ-strongly convex with high probability, for some positive $\lambda = \Theta(1)$. Putting these results together implies $\sqrt{n}$-consistency. □

2.3 Impossibility of consistent estimation

Finally, we highlight a result of a different vein. In Ghosal and Mukherjee (2020), the authors consider a special case of the general model of Daskalakis et al. (2019) in which all external fields are equal to the parameter B. The resulting model is denoted $\mathbb{P}_{n,\beta,B}(\sigma)$. We discuss an impossibility result for consistent estimation in a certain region of parameter space when the coupling matrix A is obtained by sampling an dense Erdos-Renyi random graph with edge probability $p > 0$ of constant order.

Theorem 2.6. *Given n, β, B, define the Erdos-Renyi Ising model $\mathbb{P}^{er}_{n,\beta,B}$ as a joint distribution on pairs (σ, G_n) as follows: first, sample an Erdos-Renyi random graph with adjacency matrix $G_n \sim G(n, p)$, then define the normalized matrix $A_n = \frac{1}{p(n-1)} G_n$, and finally sample a configuration $\sigma \in \{\pm 1\}^n$ from the Ising model $\mathbb{P}_{n,\beta,B}$ with coupling matrix A_n. Given $t \in (0, 1)$, define Θ_t as the set of all pairs $(B, \beta) \in \mathbb{R}^2_+$ such that $t = \tanh(\beta t + B)$. Then under $\mathbb{P}^{er}_{n,\beta,B}$, there is no consistent sequence of estimators for (B, β).*

Proof. In order to prove inconsistency, we rely on the concept of contiguity. Given a pair of sequences of measures $\mathbb{Q}_n$ and $\mathbb{P}_n$, $\mathbb{Q}_n$ is contiguous to $\mathbb{P}_n$ if for any sequence of events (E_n)

such that $\mathbb{P}_n(E_n) \to 0$, we also have $\mathbb{Q}_n(E_n) \to 0$. We show that $\mathbb{Q}_n \times G(n,p)$ is contiguous to $\mathbb{P}^{\text{er}}_{n,\beta,B}$, where $\mathbb{Q}_n$ is the product measure on $\{\pm 1\}^n$ such that $\mathbb{Q}_n(\sigma_i = +1) = \alpha$, where $\alpha = (1 + e^{-2(\beta t + B)})^{-1}$. Since $\beta t + B = \tanh^{-1} t$ by assumption, the measure $\mathbb{Q}_n \times G(n,p)$ does not depend on B and β.

In particular, if one had a $\mathbb{P}^{\text{er}}_{n,\beta,B}$-consistent estimator $(\hat{B}, \hat{\beta})$ of (B, β), then $(\hat{B}, \hat{\beta})$ would necessarily be $\mathbb{Q}_n \times G(n,p)$-consistent. But taking two distinct pairs (B_1, β_1) and (B_2, β_2), it would follow that as $n \to \infty$, all of the mass of the measure $\mathbb{Q}_n \times G(n,p)$ (which does not depend on B and β) concentrates around both (B_1, β_1) and (B_2, β_2), a contradiction.

We prove contiguity in two steps, making use of an auxiliary measure $\mathbb{P}^{\text{cw}}_{n,\beta,B}$. Here the Curie-Weiss model $\mathbb{P}^{\text{cw}}_{n,\beta,B}$ is the Ising model whose coupling matrix is given by $A_n = \frac{1}{n}\mathbf{1}\mathbf{1}^T$, corresponding to the complete graph.

First, we show that $\mathbb{P}^{\text{cw}}_{n,\beta,B}$ is contiguous to $\mathbb{Q}_n$. The proof comes down to the fact that, conditional on a certain random variable, $\mathbb{P}^{\text{cw}}_{n,\beta,B}(\sigma_i = +1)$ is very close to α.

To simplify our study of $\mathbb{P}^{\text{cw}}_{n,\beta,B}$, we make use of an auxiliary real-valued random variable W_n. Given σ, W_n is distributed as $N(\bar{\sigma}, \frac{1}{n\beta})$, where $\bar{\sigma}$ denotes the average spin. that has the following properties. The key claim is that conditional on W_n, the distribution of X is a product measure, with

$$\mathbb{P}^{\text{cw}}_{n,\beta,B}(\sigma_i = +1 | W_n) = (1 + e^{-2(\beta W_n + B)})^{-1}.$$

Further, the density of W_n is given by $\exp(-nf(w))$, where

$$f(w) = \frac{\beta w^2}{2} + Bw - \log\cosh w.$$

Looking ahead to contiguity, note that since the average spin $\bar{\sigma}$ concentrates around t, the conditional probability that $\sigma_i = +1$ is close to α.

The random variable W_n is cleverly chosen to "decouple" the spins of the Curie-Weiss model. Algebraically, the choice of conditional density cancels out cross-terms. To see the first claim, note that the Curie-Weiss model has coupling matrix $A_n = \frac{1}{n}\mathbf{1}\mathbf{1}^T$, so its probability mass function is given by

$$\begin{aligned}\mathbb{P}^{\text{cw}}_{n,\beta,B}(\sigma) &= \frac{1}{Z_n}\exp\left(\frac{\beta}{2}\sigma^T A_n \sigma + \sum_i \sigma_i B\right) \\ &= \frac{1}{Z_n}\exp\left(\frac{n\beta}{2}\bar{\sigma}^2 + nB\bar{\sigma}\right).\end{aligned}$$

By Bayes rule, the joint distribution of (σ, W_n) is given by

$$\mathbb{P}(\sigma, W_n) = \mathbb{P}(W_n|\sigma)\mathbb{P}^{\text{cw}}_{n,\beta,B}(\sigma) = \frac{1}{\sqrt{2\pi/(n\beta)}}\exp\left(-\frac{1}{2/(n\beta)}(W_n - \bar{\sigma})^2\right) \cdot \frac{1}{Z_n}\exp\left(\frac{n\beta}{2}\bar{\sigma}^2 + nB\bar{\sigma}\right).$$

Simplifying the exponential, we obtain

$$\mathbb{P}(\sigma, W_n) = \frac{1}{\sqrt{2\pi/(n\beta)}}\frac{1}{Z_n}\exp\left(-\frac{n\beta}{2}W_n^2\right) \cdot \exp((\beta W_n + B) \cdot n\bar{\sigma}).$$

Conditional on W_n, we obtain the product measure proportional to $\prod_i \exp((\beta W_n + B)\sigma_i)$, which is

$$\prod_i (1 + \exp(-2(\beta W_n + B)\sigma_i))^{-1}.$$

The second claim can be proven by marginalizing out σ, but we exclude this computation. For more details, see (Mukherjee et al., 2018, Supplementary Materials).

We compute the likelihood ratio of $\mathbb{Q}_n$ and $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}$ conditional on W_n. By definition, this equals

$$\log \frac{\mathbb{Q}_n(\sigma|W_n)}{\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(\sigma|W_n)} = S_n \log \frac{\mathbb{Q}_n(\sigma_i = +1)}{\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(\sigma_i = +1|W_n)} + (n - S_n) \log \frac{\mathbb{Q}_n(\sigma_i = -1)}{\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(\sigma_i = -1|W_n)},$$

where S_n denotes the number of positive spins in the configuration σ. One can linearize around α to obtain

$$\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(\sigma_i = +1|W_n) = \alpha + \frac{\tilde{W}_n}{\sqrt{n}},$$

where $\tilde{W}_n = \xi_n \sqrt{n}(W_n - t)$ for some bounded ξ_n. One can also show that $\tilde{W}_n = O_p(1)$ under both $\mathbb{P}^{\mathrm{cw}}_n$ and $\mathbb{Q}_n$. Plugging in, linearizing the logarithms, and rearranging, our likelihood ratio becomes

$$\begin{aligned}
\log \frac{\mathbb{Q}_n(\sigma|W_n)}{\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(\sigma|W_n)} &= S_n \log \frac{\alpha}{\alpha + \frac{\tilde{W}_n}{\sqrt{n}}} + (n - S_n) \log \frac{1 - \alpha}{1 - \alpha - \frac{\tilde{W}_n}{\sqrt{n}}} \\
&= -\frac{\tilde{W}_n}{\sqrt{n}} \frac{S_n - n\alpha}{\alpha(1 - \alpha)} + O\left(\frac{1}{n}|S_n - n\alpha|^2\right) + O(\tilde{W}_n^2).
\end{aligned}$$

Fix $K > 0$. Let $F_{n,K}$ denote the event that $|S_n - n\alpha| \leqslant \sqrt{n}$ and $|\tilde{W}_n| \leqslant K$. For $(\sigma, W_n) \in F_{n,K}$, we have the bound

$$\mathbb{Q}_n(\sigma|W_n) \leqslant \exp(z_K) \mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(\sigma|W_n),$$

where

$$z_K = \frac{K^2}{\alpha(1 - \alpha)} + O(K^2).$$

Marginalizing out W_n, we deduce that for any sequence of events (E_n),

$$\mathbb{Q}_n(E_n \cap F_{n,K}) \leqslant \exp(z_K) \mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(E_n),$$

so that

$$\mathbb{Q}_n(E_n) \leqslant \exp(z_K) \mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(E_n) + \mathbb{Q}_n(F_{n,K}).$$

By the LLN, $\mathbb{Q}_n(|S_n - n\alpha| > K\sqrt{n}) \to 0$ as $n \to \infty$ and $K \to \infty$. Since $\tilde{W}_n = O_p(1)$, $\mathbb{Q}_n(|\tilde{W}_n| > K) \to 0$ as $n \to \infty$ and $K \to \infty$. Thus $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B}(E_n) \to 0$ implies $\mathbb{Q}_n(E_n) \to 0$, implying contiguity.

Next, we show that $\mathbb{P}^{\mathrm{cw}}_{n,\beta,B} \times G(n,p)$ is contiguous to $\mathbb{P}^{\mathrm{er}}_{n,\beta,B}$, which implies the result. It suffices to show that the KL divergence between the models is of constant order (Bhattacharya

and Mukherjee, 2018, Proposition 6.1). The likelihood ratio conditional on the realization of the Erdos-Renyi graph G_n is given by

$$\begin{aligned}
&= \frac{(Z_n^{\mathrm{cw}})^{-1} \exp(\frac{\beta}{(n-1)} \sigma^T \mathbf{1}\mathbf{1}^T \sigma + nB\bar{\sigma})}{(Z_n^{\mathrm{er}})^{-1} \exp(\frac{\beta}{p(n-1)} \sigma^T G_n \sigma + nB\bar{\sigma})} \\
&= \exp\left(\sum_{i,j} \left(1 - \frac{G_n(i,j)}{p}\right) \sigma_i \sigma_j \right) \frac{Z_n^{\mathrm{er}}}{Z_n^{\mathrm{cw}}}.
\end{aligned}$$

Taking a log and an expectation over $G \sim G(n,p)$, since $\mathbb{E}G_n(i,j) = p$, the sum vanishes, and we are left with

$$\mathrm{KL}(\mathbb{P}^{\mathrm{cw}}_{n,\beta,B} \times G(n,p) || \mathbb{P}^{\mathrm{er}}_{n,\beta,B}) = \mathbb{E} \log \frac{Z_n^{\mathrm{er}}}{Z_n^{\mathrm{cw}}},$$

which by Jensen's inequality is at most $\log(\mathbb{E}Z_n^{\mathrm{er}}/Z_n^{\mathrm{cw}})$.

We now estimate the right-hand side. Rewrite the partition function Z_n^{er} as

$$Z_n^{\mathrm{er}} = \prod_{i<j} \exp\left(\frac{\beta}{n-1} 2 \frac{G_n(i,j)}{p} \sigma_i \sigma_j \right) \exp(nB\bar{\sigma}).$$

In order to bound this by Z_n^{cw}, we must replace the quantity $G_n(i,j)/p$ appearing in the first exponential by 1. By Hoeffding's lemma applied to the random variable $G_n(i,j) - p$, which takes values in $\{-p, 1-p\}$, we have

$$\mathbb{E} \exp(t(G_n(i,j) - p)) \leqslant \exp(t^2/8).$$

Plugging in $t = \frac{\beta}{n-1} 2 \frac{1}{p} \sigma_i \sigma_j$, we obtain

$$\mathbb{E} \exp\left(\frac{\beta}{n-1} 2 \frac{G_n(i,j)}{p} \sigma_i \sigma_j \right) \leqslant \exp\left(\frac{t^2}{8} \right) \exp\left(\frac{\beta}{n-1} 2 \sigma_i \sigma_j \right).$$

Multiplying over all $\binom{n}{2}$ pairs $i < j$ and recombining terms, we obtain

$$\begin{aligned}
\mathbb{E}Z_n^{\mathrm{er}} &\leqslant \exp\left(\binom{n}{2} \frac{t^2}{8} \right) Z_n^{\mathrm{cw}} \\
&\leqslant \exp\left(\frac{n(n-1)}{2} \cdot \frac{\beta^2}{(n-1)^2} 4 \frac{1}{p^2} \right) Z_n^{\mathrm{cw}} \\
&= \exp(O(1)) Z_n^{\mathrm{cw}}.
\end{aligned}$$

Thus the KL divergence is bounded. □

3 Follow-up work

The paper Kandiros et al. (2021) extends the results of Daskalakis et al. (2019) to a broader class of external fields. Given a finite-dimensional parametric function class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$,

consider the model $\mathbb{P}_{n,\beta,\theta}$ whose external field at site i is given by $f_\theta(x_i)$. Here the external field is modeled as a potentially nonlinear function f_θ of the features $x \in \mathbb{R}^d$. This paper also relaxes some assumptions from Daskalakis et al. (2019): for example, they do not impose the bound $\|A_n\|_F^2 \geqslant cn$, but instead adapt the estimation rate to smaller values of $\|A_n\|_F^2$.

In order to state the main result (Kandiros et al., 2021, Theorem 6), we must make the following definition. Given a function class $\mathcal{F}$ and a distance metric $d(\cdot,\cdot)$ on $\mathcal{F}$, an ε-net is a subset of $\mathcal{F}$ such that for any function $f \in \mathcal{F}$, there exists a function g in the ε-net such that $d(f,g) \leqslant \varepsilon$. The size of the smallest ε-net is denoted $N(\mathcal{F}, d, \varepsilon)$ and is called the ε-covering number of the metric space $(\mathcal{F}, d)$.

Theorem 3.1. *Suppose that $\beta \in [-\beta, \beta]$, $\theta \in \Theta$, $\|A_n\|_\infty = 1$, and $|f_\theta(x_i)| \leqslant M$ uniformly over $\theta \in \Theta$ and $i \in [n]$ for some constant M. Define the distance on $\mathcal{F}$ given by*

$$d(f, f') = \|f - f'\|_2^2 = \frac{1}{n}\sum_i (f(x_i) - f'(x_i))^2.$$

Then with probability at least $1-\delta$, given a sample $\sigma \sim \mathbb{P}_{n,\beta,\theta}$, the MPLE $(\hat{\theta}_n, \hat{\beta}_n)$ obeys

$$\|f_{\hat{\theta}_n} - f_\theta\|_2^2 \leqslant C_1 \inf_{\varepsilon \geqslant 0} \left(\log \frac{n}{\delta} + \varepsilon n + \log N(\mathcal{F}, d, \varepsilon) \right)$$

and

$$\|\hat{\beta}_n - \beta\|_2^2 \leqslant C_2 \inf_{\varepsilon \geqslant 0} \left(\log \frac{n}{\delta} + \varepsilon n + \log N(\mathcal{F}, d, \varepsilon) \right),$$

where the constants C_1, C_2 are at most $1/\|A\|_F^2$.

As a special case of this general result, by considering the linear function class $f_\theta(x) = x^T\theta$ defined for $x \in \mathbb{R}^d$, Kandiros et al. (2021) obtain the following generalization of Theorem 2.4. This result, (Kandiros et al., 2021, Theorem 1), makes explicit the rate of convergence of the MPLE on $\|A_n\|_F^2$.

Theorem 3.2. *Under the same assumptions as Theorem 3.1, assuming $\|x_i\|_2 \leqslant M$ for all $i \in [n]$, and assuming $\|\theta\|_2 \leqslant 1$, then when $f_\theta(x) = x^T\theta$, given a sample $\sigma \sim \mathbb{P}_{n,\beta,\theta}$, the MPLE satisfies*

$$\|(\hat{\theta}_n, \hat{\beta}_n) - (\theta, \beta)\|_2 \leqslant C \frac{d \log n + \log(1/\delta)}{\|A_n\|_F^2}$$

with probability at least $1-\delta$.

Proof. (Sketch.) For the purpose of estimation, note that we may identify f_θ with the vector $f_\theta(X) \in \mathbb{R}^d$, the vector of external fields. The set $\mathcal{H} = \{f_\theta(X) : \theta \in \Theta\} \subseteq \mathbb{R}^n$ equipped with the metric $d(h, h') = \|h - h'\|_2^2/n$ is isometric to $(\mathcal{F}, d)$. In what follows, we pass between $(\mathcal{F}, d)$ and $(\mathcal{H}, d)$ based on convenience.

In order to handle arbitrary classes $\mathcal{F}$, Kandiros et al. (2021) first study the performance of $\hat{\theta}$ in a one-dimensional class. In the case that $\mathcal{H} = h + \operatorname{span}\{v\} \subseteq \mathbb{R}^n$ for some v, a strong convexity-based analysis similar to Daskalakis et al. (2019) yields

$$\|f_{\hat{\theta}} - f_\theta\|_2^2 \leqslant C \frac{\log \frac{n}{\delta}}{\|A_n\|_F^2}$$

with probability at least $1-\delta$. Given N directions $v_1, \dots, v_N$ in $\mathbb{R}^n$, we can compute the MPLE $\hat{\theta}_i$ along each individual one-dimensional affine subspace $h + \operatorname{span}\{v_i\}$, where $h = f_\theta(X)$ is the true external field vector. Then by a union bound, with probability at least $1-\delta$, for all $i \in [N]$ we have

$$\|f_{\hat{\theta}_i} - f_\theta\|_2^2 \leqslant C \log \frac{n}{\delta/N} = C\left(\log \frac{n}{\delta} + \log N\right)$$

The key observation is that if the MPLE $\hat{\theta}$ computed over the full function class lies in the union $\bigcup_{i=1}^N (h + \operatorname{span}\{v_i\})$, then necessarily $\hat{\theta} = \hat{\theta}_i$ for some $i \in [N]$, so that the bound holds.

In order to ensure that $\hat{\theta}$ is within ε of the union, we let $\{v_1, \dots, v_N\}$ form an ε-net for $\mathcal{H}$, where $N = N(\mathcal{H}, d, \varepsilon)$ (which, by isometry, is also $N(\mathcal{F}, d, \varepsilon)$). Since the estimation error is Lipschitz on a bounded region, with Lipschitz constant proportional to the number of terms in the sum, the estimation error $\|f_{\hat{\theta}} - f_\theta\|_2^2$ of the MPLE is within $O(\varepsilon n)$ of the estimation error restricted to $\bigcup_{i=1}^N (h + \operatorname{span}\{v_i\})$. Putting this all together, we deduce

$$\|f_{\hat{\theta}} - f_\theta\|_2^2 \leqslant C\left(\log \frac{n}{\delta} + \varepsilon n + \log N(\mathcal{F}, d, \varepsilon)\right),$$

and the result follows by optimizing over $\varepsilon \geqslant 0$. □

References

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

B. B. Bhattacharya and S. Mukherjee. Inference in ising models. *Bernoulli*, 24(1):493–525, 2018.

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

S. Chatterjee. Estimation in spin glasses: a first step. *Annals of Statistics*, 35(5):1931–1946, 2007.

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

P. Ghosal and S. Mukherjee. Joint estimation of parameters in ising model. *The Annals of Statistics*, 48(2):785–810, 2020.

V. Kandiros, Y. Dagan, N. Dikkala, S. Goel, and C. Daskalakis. Statistical estimation from dependent data. In *International Conference on Machine Learning*, pages 5269–5278. PMLR, 2021.

R. Mukherjee, S. Mukherjee, and M. Yuan. Global testing against sparse alternatives under ising models. *The Annals of Statistics*, 46(5):2062–2093, 2018.