Statistical Estimation from Dependent Data

February 12, 2021

Abstract

We consider a general statistical estimation problem wherein binary labels across different observations are not independent conditioned on their feature vectors, but dependent, capturing settings where e.g. these observations are collected on a spatial domain, a temporal domain, or a social network, which induce dependencies. We model these dependencies in the language of Markov Random Fields and, importantly, allow these dependencies to be substantial, i.e. do not assume that the Markov Random Field capturing these dependencies is in high temperature. As our main contribution we provide algorithms and statistically efficient estimation rates for this model, giving several instantiations of our bounds in logistic regression, sparse logistic regression, and neural network settings with dependent data. Our estimation guarantees follow from novel results for estimating the parameters (i.e. external fields and interaction strengths) of Ising models from a *single* sample. We evaluate our estimation approach on real networked data, showing that it outperforms standard regression approaches that ignore dependencies, across three text classification datasets: Cora, Citeseer and Pubmed.

1 Introduction

The standard supervised learning framework assumes access to a collection $(x_i, y_i)_{i=1}^n$ of observations, where the labels $y_1, \ldots, y_n \in \mathcal{Y}$ are independent conditioning on the feature vectors $x_1, \ldots, x_n \in \mathcal{X}$. Further, it is common to assume that each label y_i is independent of $\{x_j\}_{j \neq i}$ conditioning on x_i, i.e. that

$$\mathbb{P}[y_{1\ldots n} \mid x_{1\ldots n}] = \prod_{i=1}^{n} \mathbb{P}[y_i \mid x_i],$$

and, moreover, that the observations share the same generative process $\mathbb{P}[y \mid x]$ sampling a label conditioning on a feature vector. Under these assumptions, a common goal is to identify a model $\mathbb{P}_\theta[y \mid x]$ from some parametric class, which approximates the true generative process $\mathbb{P}[y \mid x]$ in some precise sense, or, under realizability assumptions, to estimate the parameter θ of the true generative process. A special case of this problem is the familiar logistic regression problem, where each label lies in $\mathcal{Y} = \{\pm 1\}$, each feature vector lies in $\mathbb{R}^d$ and for some $\theta \in \mathbb{R}^d$ it is assumed that

$$\mathbb{P}[y_{1\ldots n} \mid x_{1\ldots n}] = \prod_{i=1}^{n} \frac{1}{1 + \exp(-2(\theta^\top x_i) y_i)}. \tag{1}$$

The standard assumptions outlined above are, however, too strong and almost never truly hold in practice. Indeed, they become especially prominent when it comes to observations collected in a temporal domain, a spatial domain or a social network, which naturally induce dependencies among the observations. Such dependencies could arise from physical constraints, causal relationships among observations, or peer effects in a social network. They have been studied extensively in many practical fields, and from a theoretical standpoint in econometrics and statistical learning theory. See section 1.2 for further discussion.

In this paper we study such dependencies conforming to the following general class of models:

$$\mathbb{P}[y_{1\ldots n} \mid x_{1\ldots n}] \propto \exp(-\beta \cdot H(\vec{y})) \cdot \prod_{i=1}^{n} \exp(f_\theta(x_i, y_i)) \equiv \exp\left(-\beta \cdot H(\vec{y}) + \sum_{i=1}^{n} f_\theta(x_i, y_i)\right), \tag{2}$$

where f_θ is an (unknown) function from some parametric class, H is a (known) function that captures the dependency structure and β is an (unknown) parameter that captures the strengths of dependencies. It should be appreciated that Model (2) is more general than the standard supervised learning problem without dependencies, which results from setting $\beta = 0$. Once we allow $\beta \neq 0$, Model (2) becomes more expressive in capturing the dependencies among the observations, which become stronger with higher values of β. The challenging estimation problem that arises, which motivates our work, is whether the model parameters θ and/or β can be identified, and at what rates, in the presence of the intricate dependencies arising from this model. Importantly, while the labels are intricately dependent, we do not have access to multiple independent samples from the conditional distribution (2), but a *single* sample from that distribution!

We focus here on a special case of Model (2) wherein the labels are binary and the function H is pairwise separable, studying models of the following form:

$$\mathbb{P}_{\theta,\beta}[y_{1\dots n} \mid x_{1\dots n}] = \frac{\exp(\beta y^{\mathrm{T}} A y) \prod_{i=1}^{n} \exp(y_i f_\theta(x_i))}{Z_{\theta,\beta}} \equiv \frac{1}{Z_{\theta,\beta}} \exp\left(\beta y^{\mathrm{T}} A y + \sum_i y_i f_\theta(x_i)\right), \tag{3}$$

where f_θ is an unknown function from some parametric class $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathbb{R} \mid \theta \in \Theta\}$, A is a known, symmetric *interaction matrix* with zeros on the diagonal, β is an unknown parameter, and $Z_{\theta,\beta}$ is the normalizing constant. In other words, under (3), the labels $y_1, \dots, y_n$ are sampled from an n-variable *Ising model* with *external field* $f_\theta(x_i)$ on variable i, *interaction strength* $A_{ij} \equiv A_{ji}$ between variables i and j, and *inverse temperature* β. Notice that A_{ij} encourages y_i and y_j to have the same or opposite values, depending on its sign, however, this "local encouragement" can be overwritten by indirect interactions through other values of y_k. Such indirect interactions make this model rich in spite of the simple form of $H(y) = y^\top A y$ and as a consequence, it has found profound applications in a range of disciplines, including Statistical Physics, Computer Vision, Computational Biology, and the Social Sciences; see e.g. [GG86; Ell93; Fel04; Cha05; DMR11; DDK17].

It is clear that Model (3) generalizes (1), which can be obtained by setting $\beta = 0$, $\mathcal{X} = \mathbb{R}^d$, and $f_\theta(x) = \theta^{\mathrm{T}} x$. It also generalizes the model studied by Daskalakis et al. [DDP19], which results from setting $f_\theta(x) = \theta^{\mathrm{T}} x$ and $0 \leq \beta \leq O(1)$, as well as the model studied by Ghosal and Mukherjee [GM18], Bhattacharya and Mukherjee [BM18], and Chatterjee [Cha07], which results from taking $f_\theta(x)$ to be a constant function.

We study under what conditions on the function class $\mathcal{F}$, the interaction matrix A, and the feature vectors $(x_i)_i$, and at what rates can the parameters θ and/or β of Model (3) be estimated given a collection $(x_i, y_i)_{i=1}^n$ of observations, where the labels $y_1, \dots, y_n$ are sampled from (3) conditioning on the feature vectors $x_1, \dots, x_n$. As explained earlier, in comparison to the standard supervised learning setting without dependencies, the statistical challenge that arises here is that, while the labels $y_1, \dots, y_n$ are intricately dependent, we do not have access to multiple independent samples from the conditional distribution (3), but a single sample from that distribution. Thus, it is not clear how to extract good estimates of the parameters from our observations and where to find statistical power to bound the error of these estimates from the true parameters. As a consequence, only limited theoretical results in this area are known.

1.1 Overview of Results

We provide a general algorithmic approach which yields efficient statistical rates for the estimation of θ and/or β of Model (3) for general function classes $\mathcal{F}$, in terms of the metric entropy of $\mathcal{F}$. We also prove information theoretic lower bounds, which combined with our upper bounds characterize the min-max estimation rate of the problem up to a certain factor, discussed below. Before stating our general result as Theorem 6, we present some corollaries of this theorem in more familiar settings. All the theorems that follow are also presented and proved in more detail in the Supplementary Material. Finally, in all statements below we use the following notation and assumptions, which summarize the already described setting.

Assumptions 1 (and useful Notation)**.** We are given observations $(x_i, y_i)_{i=1}^n$, where $y_1, \dots, y_n$ are sampled from (3) conditioning on $x_1, \dots, x_n$, using some unknown parameters $\theta^* \in \Theta$ and $\beta^* \in [-B, B]$, and some known A, normalized such that $\|A\|_\infty = 1$. We further assume that $|f_\theta(x_i)| \leq M$, for all i and $\theta \in \Theta$. In all

statements below, $\hat{\theta}$ and $\hat{\beta}$ refer to the estimates produced by the algorithm described in Section 2, i.e. the Maximum Pseudo-Likelihood Estimator (MPLE). Moreover, we let $\lesssim$ denote an inequality up to factors that are singly-exponential in M and B, a necessary dependence on those parameters when $\lesssim$ is used, and are independent of all other parameters. In particular, when $M, B = O(1)$, $\lesssim$ denotes inequality up to a constant.

Under the assumptions on our observations, and notation introduced above, we consider two settings to illustrate our general result (Theorem 6), namely linear classes (Setting 1) and neural network classes (Setting 2).

Setting 1 (Linear Classes)**.** Make Assumptions 1, suppose $x_i \in \mathbb{R}^d$ and $\|x_i\|_2 \leq M$, for all i, and suppose that f_θ is linear, i.e. $f_\theta(x_i) = x_i^\top \theta$, for some $\theta \in \mathbb{R}^d$ and $\|\theta\|_2 \leq 1$. Denote by X the matrix whose rows are $x_1, \ldots, x_n$ and by κ the minimum eigenvalue of $\frac{X^T X}{n}$, or its minimum restricted eigenvalue in the sparse setting of Theorem 2. We suppress from our bounds of Theorems 1 and 2 a factor of $1/\kappa$.

Theorem 1 (Linear Class)**.** *Suppose Setting 1. Then, with probability* $\geq 1 - \delta$*,*

$$\|\hat{\theta} - \theta^*\|_2^2 + |\hat{\beta} - \beta^*|^2 \lesssim \frac{d \log n + \log(1/\delta)}{\|A\|_F^2}.$$

Theorem 2 (Sparse Linear Class)**.** *Suppose Setting 1 and additionally that* $\|\theta\|_1 \leq s$*. Then, w.pr.* $\geq 1 - \delta$*,*

$$\|\hat{\theta} - \theta^*\|_2^2 + |\hat{\beta} - \beta^*|^2 \lesssim \frac{(n^2 s \log(d))^{1/3} + \log(1/\delta)}{\|A\|_F^2}.$$

Theorem 1 is proved in Section D.1 and Theorem 2 is proved in Section D.2.

Both bounds above are obtained by minimizing a convex function over a convex domain, which can be performed in polynomial time. We note that the bound of Theorem 1 generalizes the main result of Daskalakis et al. [DDP19], which makes the additional assumption that $\|A\|_F = \Omega(\sqrt{n})$. We need no such assumption and our bound gracefully degrades as $\|A\|_F$ decreases. Theorem 2 extends these results to the sparse linear model, for which no prior results exist. Note that our bound is non-vacuous as long as $\|A\|_F = \Omega(n^{1/3})$, which is a reasonable expectation, given that A is $n \times n$. Moreover, it is possible to remove the appearance of n^2 from the bound of this theorem, if our model class satisfies $|\theta|_0 \leq s$. Finally, we note that the factor $1/\|A\|_F^2$ which appears in our error bounds is tight, as per the following.

Theorem 3 (Lower bound)**.** *For any* n *and* $r \in [1, n]$ *there exists an instance of a* $d = 1$*-dimensional linear class that satisfies the assumptions of Theorems 1 and 2 and further* $\|A\|_F^2 = r$*, such that any estimator* (θ', β') *satisfies with probability* ≥ 0.49*,*

$$|\theta' - \theta^*|^2 \gtrsim \frac{1}{\|A\|_F^2}, \quad |\beta' - \beta^*|^2 \gtrsim \frac{1}{\|A\|_F^2}.$$

Theorem 3 is proved in Section C.2. While Theorem 3 shows that a dependence in $\frac{1}{\|A\|_F^2}$ is unavoidable in the worst case, under favorable assumptions we can remove such dependence as per the following theorem.

Theorem 4 (Linear Class, Random Features)**.** *In the same setting as Theorem 1, remove all assumptions involving the feature vectors and suppose instead that* $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$*. Then, with probability* $\geq 1 - \delta$*,*

$$\|\hat{\theta} - \theta^*\|_2^2 \lesssim \xi(n, 1/\delta) \frac{d + \log(1/\delta)}{n} \left(1 + \frac{d + \log(1/\delta)}{\|A\|_F^2 / \|A\|_2^2}\right),$$

where $\xi(n, \frac{1}{\delta})$ *is linear in* $\log\log(\frac{1}{\delta})$ *and sub-polynomial (i.e. asymptotically smaller than any polynomial) in* n*.*

Theorem 4 is proved in Section D.1. Noticing that $\|A\|_F^2/\|A\|_2^2 \geq 1$, Theorem 4 shows that no lower bound on $\|A\|_F$ is necessary at all, if we are only looking to estimate θ^*, which answers a main problem left open by Daskalakis et al. [DDP19]. Moreover, when $\|A\|_F^2/\|A\|_2^2 \geq d$, which is a reasonable expectation in our setting since $\|A\|_2 \leq 1$ and A is $n \times n$, our bound here essentially matches the estimation rates known for the familiar logistic regression problem, which corresponds to the case $\beta = 0$, even though we make no such assumption, and hence our labels are dependent.

Beyond linear and sparse linear function classes, our main result (Theorem 6) provides estimation rates for neural network regression, as in the following setting.

Setting 2 (Neural Networks)**.** Make Assumptions 1 and suppose that the function f_θ in (3) is a neural network parameterized by θ. We adopt the setting and terminology of [BFT17]. In particular, we assume that the neural network takes the form:

$$f_\theta(x) = \sigma_L(W_L\sigma_{L-1}(W_{L-1}\cdots\sigma_1(W_1x)\cdots)), \tag{4}$$

where the depth L of the network is fixed, $\sigma_1, \ldots, \sigma_L : \mathbb{R} \to \mathbb{R}$ are some fixed non-linearities, and $W_1, \ldots, W_L$ are (unknown) weight matrices. In particular, $\theta = (W_1, \ldots, W_L)$.

We denote by $\rho_1, \ldots, \rho_L$ the Lipschitz constants of the non-linearities, and when, abusing notation, we apply some non-linearity σ_i to a vector v, the result $\sigma_i(v)$ is a vector whose j-th coordinate is $\sigma_i(v_j)$. We also adopt from [BFT17] the notion of *spectral complexity* R_θ of a neural network f_θ with respect to reference matrices $M_1, \ldots, M_L$ (of the same dimensions as $W_1, \ldots, W_L$ respectively), defined in terms of different matrix norms as follows:

$$R_\theta = \left(\prod_{i=1}^{L} \rho_i \|W_i\|_2\right)\left(\sum_{i=1}^{L} \frac{\|W_i^{\mathrm{T}} - M_i^{\mathrm{T}}\|_{2,1}^{2/3}}{\|W_i\|_2^{2/3}}\right)^{3/2},$$

where $\|M\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{n} M_{ij}^2}$. Assuming a fixed bound on each matrix norm involved in the above expression, we take $\mathcal{F} = \{f_\theta \colon \theta \in \Theta\}$ to be the collection of all neural networks of Form (4), whose weight matrices satisfy those bounds. Suppose R is the resulting bound on the spectral norm of all networks in our family, implied by our assumed bounds on the various matrix norms. Finally, we assume that the widths of all networks $f_\theta \in \mathcal{F}$ are bounded by d.

Theorem 5. *Suppose Setting 2, and let* $K^2 = \frac{1}{n}\sum_i \|x_i\|_2^2$*. Then, with probability* $\geq 1 - \delta$*,*

$$\frac{1}{n}\sum_{i=1}^{n}(f_{\hat{\theta}}(x_i) - f_{\theta^*}(x_i))^2 + |\hat{\beta} - \beta^*|^2 \quad \lesssim \frac{(n^2K^2R^2\log d)^{1/3} + \log\left(\frac{n}{\delta}\right)}{\|A\|_F^2}.$$

Theorem 5 is proved in Section D.3. Notice that, in this case, we do not provide guarantees for the estimation of θ. Since these networks are often overparametrized, it might be impossible to recover θ.

All estimation results above, namely Theorems 1–5, are corollaries of our general estimation result given below.

Theorem 6 (General Estimation Result)**.** *Make Assumptions 1, where* f_{θ^*} *lies in some general class* $\mathcal{F} = \{f_\theta\}_\theta$*. Then, w.pr.* $\geq 1 - \delta$*,*

$$\frac{1}{n}\sum_{i=1}^{n}(f_{\hat{\theta}}(x_i) - f_{\theta^*}(x_i))^2 \lesssim \mathcal{C}_1(\mathcal{F}, X, \beta^*, \theta^*) \inf_{\epsilon \geq 0}\left(\log\frac{n}{\delta} + \epsilon n + \log N(\mathcal{F}, X, \epsilon)\right),$$

where X *denotes the collection of feature vectors,* $N(\mathcal{F}, X, \epsilon)$ *is the* ϵ*-covering number of* $\mathcal{F}$ *under distance* $d(f, f') = \sqrt{\sum_{i=1}^{n}(f(x_i) - f'(x_i))^2/n}$ *and* $\mathcal{C}_1 \leq 1/\|A\|_F^2$ *is a quantity that has a simple formula (both quantities are formally defined in Section 3.2). Further, if* $\mathcal{F}$ *is convex and closed under negation,*[1] *for any*

[1] We say that $\mathcal{F}$ is convex if for any $f, f' \in \mathcal{F}$ and any $\lambda \in [0, 1]$ the function $\tilde{f}(x) = (1 - \lambda)f(x) + \lambda f'(x)$ belongs to $\mathcal{F}$. We say that $\mathcal{F}$ is closed under negation if $-f \in \mathcal{F}$ for all $f \in \mathcal{F}$.

estimator (θ', β') *there exists* (θ^*, β^*)*, s.t. w.pr.* ≥ 0.49*,*

$$\frac{1}{n}\sum_{i=1}^{n}(f_{\theta'}(x_i) - f_{\theta^*}(x_i))^2 \gtrsim \mathcal{C}_1(\mathcal{F}, X, \beta^*, \theta^*).$$

Similar upper and lower bounds hold for estimating β^**, with* $\mathcal{C}_1$ *replaced with a different quantity* $\mathcal{C}_2 \leq 1/\|A\|_F^2$*.*

Theorem 6 is proved in Section A. Theorem 6 is used to derive Theorems 1, 2, 4, and 5 by bounding the covering numbers of linear, sparse linear and neural network classes. It is also used to derive Theorem 3 in a straight-forward way. It is worth emphasizing that we obtain separate general estimation rates for β and θ, which are tight or near-tight in a variety of settings.

1.2 Related Work

Data dependencies are pervasive in many applications of Statistics and Machine Learning, e.g. in financial, meteorological, epidemiological, and geographical applications, as well as social-network analyses, where peer effects have been studied in topics as diverse as criminal activity [GSS96], welfare participation [BLM00], school achievement [Sac01], retirement plan participation [DS03], and obesity [CF13; TNP08]. These applications have motivated substantial work in Econometrics (see e.g. Manski [Man93] and Bramoullé et al. [BDF09] and their references), where identification results have been pursued and debated, mostly in linear auto-regressive models; see also Daskalakis et al. [DDP19]. In Statistical Learning Theory, learnability and uniform convergence bounds have been shown in the presence of sample dependencies; see e.g. Yu [Yu94], Gamarnik [Gam03], Berti et al. [Ber+09], Mohri and Rostamizadeh [MR09], Pestov [Pes10], Mohri and Rostamizadeh [MR10], Shalizi and Kontorovich [SK13], London et al. [Lon+13], Kuznetsov and Mohri [KM15], London et al. [LHG16], McDonald and Shalizi [MS17], and Dagan et al. [Dag+19]. Those learnability frameworks are not applicable to our setting due to exchangeability, fast-mixing, or weak-dependence properties that they are exploiting.

Close to our setting, recent work of Daskalakis et al. [DDP19] considers a special case of our problem, where function f_θ in Model (3) is assumed linear. We obtain stronger estimation bounds, under weaker assumptions, our bounds gracefully degrading with $\|A\|_F$, as we have already discussed. Similarly, earlier work by Chatterjee [Cha07], Bhattacharya and Mukherjee [BM18], Ghosal and Mukherjee [GM18], and Dagan et al. [Dag+20], motivated by single-sample estimation of Ising models, considers a special case of our problem where function f_θ in Model (3) is assumed constant. Our bounds in this simple setting are as tight as the tightest bounds in that line of work. Overall, in comparison to these works, our general estimation result (Theorem 6) covers arbitrary classes $\mathcal{F}$, characterizing the estimation rate up to a factor that depends on the metric entropy of $\mathcal{F}$. We thus obtain rates for sparse linear classes (Theorem 2), neural network classes (Theorem 5), and Lipschitz classes (discussed in the Supplementary Material), which had not been shown before. Finally, our bounds disentangle our ability of estimating θ and β, allowing for the estimation of θ even when the estimation of β is impossible, as shown in Theorem 4 for linear classes, answering a main open problem left open by [DDP19].

At a higher level, single-sample statistical estimation is both a classical and an emerging field [Bes74; BN18; CVV19; Dag+20] with intimate connections to Statistical Physics, Combinatorics, and High-Dimensional Probability.

Roadmap. We present the estimator used to derive all our upper bounds in Section 2. We present a sketch of our proof of Theorem 6 in Section 3. We do this in two steps. First we present a sketch for the toy case of Theorem 1, i.e. the single-dimensional case. This illustrates some of the main ideas of the proof. We then provide the modifications necessary for the multi-dimensional case, which naturally lead us to the formulation of Theorem 6. While the main technical ideas are already illustrated in Section 3 in sufficient detail, the complete details can be found in the supplementary material. We conclude with experiments in Section 4, where we apply our estimator on citation datasets and compare its prediction accuracy to supervised learning approaches that do not take into account label dependencies.

2 The Estimation Algorithm

In all our theorems, the estimator we use is the Maximum Pseudo-Likelihood Estimator (MPLE), first proposed by Besag [Bes74] and defined as follows

$$(\hat{\theta},\hat{\beta}) := \arg\max_{\theta,\beta} \prod_{i=1}^{n} \mathbb{P}_{\theta,\beta}[y_i|x,y_{-i}] \equiv \prod_{i=1}^{n} \frac{\exp\left(y_i\left(f_\theta(x_i)+\beta\sum_{j=1}^{n}A_{ij}y_j\right)\right)}{2\cosh\left(f_\theta(x_i)+\beta\sum_{j=1}^{n}A_{ij}y_j\right)}\ , \tag{5}$$

where $\mathbb{P}_{\theta,\beta}$ is defined in (3), $x=(x_1,\ldots,x_n)$ and $y_{-i}=(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_n)$. We optimize the MPLE over $\theta\in\Theta$ and $\beta\in[-B,B]$, for Θ,B as per Assumption 1.

In comparison to MPLE, the more common Maximum Likelihood Estimator (MLE) optimizes $\mathbb{P}_{\theta,\beta}[y_{1\ldots n}\mid x_{1\ldots n}]$. Notice that the MPLE coincides with the MLE in the case $\beta=0$, which corresponds to $y_1,\ldots,y_n$ being independent conditioned on $x_{1\ldots n}$. When $\beta\neq 0$, this conditional independence ceases to hold and the two methods target different objectives. In this case, the objective function of MLE, which is (3), involves the normalizing factor $Z_{\theta,\beta}$, which is in general computationally hard to approximate [SS14]. In contrast, the MPLE is efficiently computable in many cases. For example, in the linear case where $f_\theta(x_i)=x_i^\top\theta$, the logarithm of (5) is a convex function of θ and β. Hence, we can use a variety of convex optimization algorithms to find the optimal solution. Even in cases where it is not a convex function, we can always use generic optimization techniques such as gradient-based methods to find a local optimum fast, since the derivative is easy to compute. Thus, the MPLE is a very appealing choice for various models. In all the results that follow, both theoretical and practical, the algorithm used will be the MPLE.

3 Proof overview

In this section, we will briefly describe the most important contributions of this work at the technical level. We start by discussing the case where f_θ is linear and θ is a one-dimensional parameter. We describe in detail the obstacles that had to be overcome to obtain tight rates for the estimation of θ and β in this case and highlight some of the most important features of the proof. In particular, we use the *mean field approximation*, a tool from statistical physics, to derive the bounds. Later, we sketch the proof of the general Theorem 6.

Notation: Matrix Norms. We use the Forbenius norm $\|A\|_F$, the spectral norm $\|A\|_2$ and the infinity-to-infinity norm $\|A\|_\infty$. In our setting A is symmetric, so one has $\|A\|_2\le\|A\|_\infty=1$ and $\|A\|_F\le\sqrt{n}\|A\|_2\le\sqrt{n}$.

3.1 Single-dimensional linear classes

We consider the setting of Theorem 1, when the dimension is $d=1$. We denote $x=(x_1,\ldots x_n)$ and $y=(y_1,\ldots,y_n)$. To simplify the presentation, we assume $\kappa\ge\Omega(1)$, which implies that $\|x\|_2\ge\Omega(\sqrt{n})$, and further that $M,B=O(1)$. In this sketch we focus on estimating θ while the bound on β is similarly obtained, and our goal is to show the special case of Theorem 1 for dimension $d=1$, namely, that with probability $\ge 1-\delta$:

$$|\hat{\theta}-\theta^*|\lesssim\frac{\sqrt{\log\frac{n}{\delta}}}{\|A\|_F}. \tag{6}$$

In fact, we will show the tighter bound of:

$$|\hat{\theta}-\theta^*|\lesssim\sup_{\lambda\in\mathbb{R}}\frac{\sqrt{\log\frac{n}{\delta}}}{\|\lambda A\|_F+\left\|x-\lambda A\mathbf{tanh}\left(\frac{\beta^* x}{\lambda}+\theta^* x\right)\right\|_2} \tag{7}$$

where $\mathbf{tanh}(z_1,\ldots,z_n)=(\tanh(z_1),\ldots,\tanh(z_n))$. We note that this bound is tight up to the factor of $\sqrt{\log\frac{n}{\delta}}$ (after a small tweak to these bounds that we omit for simplicity), and it can be obtained from our general bound of Theorem 6 with respect to the quantity $\mathcal{C}_1$ (see Section 3.2.

Before establishing (7), we note that it is stronger than the right hand side of (6). This follows from a simple exercise, considering cases for λ and utilizing the fact that under the assumptions stated above, $\|\lambda A\mathbf{tanh}((\beta^*/\lambda)x + \theta^* x)\|_2 \le O(\lambda\sqrt{n})$, while $\|x\|_2 \ge \Omega(\sqrt{n})$.

We proceed with sketching the proof of (7). Let $\varphi(\theta, \beta)$ be the negative pseudo log-likelihood for the pair (θ, β), namely, minus the log of the quantity in (5). This is a convex function whose minimum equals $(\hat{\theta}, \hat{\beta})$ and our goal is to show that (θ^*, β^*) lies in proximity to this minimum. In order to show this, it suffices to prove that the gradient of φ at (θ^*, β^*) is small, while the function is strongly convex in its neighborhood. For a more rigorous proof, we write $\varphi(\hat{\theta}, \hat{\beta})$ using a Taylor sum around (θ^*, β^*). Denoting $v = (v_\theta, v_\beta) = (\hat{\theta} - \theta^*, \hat{\beta} - \beta^*)$, we get:

$$\varphi(\hat{\theta}, \hat{\beta}) = \varphi(\theta^*, \beta^*) + v^\top \nabla\varphi(\theta^*, \beta^*) + \frac{1}{2} v^\top \nabla^2 \varphi(\theta', \beta') v,$$

for some (θ', β') in the segment connecting (θ^*, β^*) and $(\hat{\theta}, \hat{\beta})$. Since $(\hat{\theta}, \hat{\beta})$ is the minimizer of the MPLE, one has $\varphi(\hat{\theta}, \hat{\beta}) \le \varphi(\theta^*, \beta^*)$, which implies that

$$\frac{1}{2} v^\top \nabla^2 \varphi(\theta', \beta') v \le -v^\top \nabla\varphi(\theta^*, \beta^*) \le |v^\top \nabla\varphi(\theta^*, \beta^*)|. \tag{8}$$

Using concentration inequalities from [Dag+20], we can show that w.pr. $\ge 1 - \delta$ (w.r.t. the randomness of the $y_1, \dots, y_n$ which are implicit arguments of φ), any $u \in \mathbb{R}^2$ satisfies

$$\frac{|u^\top \nabla\varphi(\theta^*, \beta^*)|}{u^\top \nabla^2 \varphi(\theta', \beta') u} \lesssim \frac{\sqrt{\log n/\delta}}{\|u_\beta A\|_F + \|u_\theta x + u_\beta A y\|}. \tag{9}$$

After substituting $u = v$, it follows from (8) that the left hand side of (9) is lower bounded by $1/2$. We derive that

$$1 \lesssim \frac{\sqrt{\log n/\delta}}{\|v_\beta A\|_F + \|v_\theta x + v_\beta A y\|}.$$

Multiplying by v_θ, and writing $\lambda = -v_\beta/v_\theta$, we have

$$|\hat{\theta} - \theta^*| = |v_\theta| \le \frac{\sqrt{\log n/\delta}}{\|\lambda A\|_F + \|x - \lambda A y\|} \le \sup_{\lambda \in \mathbb{R}} \frac{\sqrt{\log n/\delta}}{\|\lambda A\|_F + \|x - \lambda A y\|}. \tag{10}$$

At this point, we have bounded the rate by the solution to an optimization problem. However, notice that the right hand side contains y which is a random variable. We would like to show that the whole expression is bounded by a nonrandom quantity and, in particular, by (7). This statement requires new insights and, as a result, a significant part of the proof is devoted to it. Here, we first sketch the main idea and then give a more technical explanation for it.

We would like to bound the optimization problem in (10) by that in (7), which corresponds to showing

$$\|\lambda A\|_F + \|x - \lambda A y\| \gtrsim \|x - \lambda A \mathbf{tanh}((\beta^*/\lambda)x + \theta^* x)\|. \tag{11}$$

We start by describing a rough and informal intuition for proving (11), and later proceed with a more formal derivation. We use an approach from statistical physics that is called *mean-field approximation*: we can substitute each y_i with $\mathbb{E}[y_i \mid x, y_{-i}] = \tanh(\beta^* \sum_j A_{ij} y_j + \theta^* x)$. Applying this substitution for all i, we obtain that

$$y \approx \mathbf{tanh}(\beta^* A y + \theta^* x). \tag{12}$$

We assume towards contradiction that (11) does not hold, and in this case we make the (false) substitution $\|x - \lambda A y\|_2 \approx 0$, which implies that $Ay \approx x/\lambda$. Substituting this in the right hand side of (12), we obtain that $y \approx \mathbf{tanh}(\beta^* x/\lambda + \theta^* x)$. Making this substitution in $\|x - \lambda A y\|$, we obtain (11).

Now, we will argue more formally about the previous claims to derive (11). Using the triangle inequality, we get

$$\|x - \lambda A\mathbf{tanh}(\beta^* x/\lambda + \theta^* x)\| \leq +\|x - \lambda Ay\| + \|\lambda Ay - \lambda A\mathbf{tanh}(\beta^* Ay + \theta^* x)\| \\ + \|\lambda A\mathbf{tanh}(\beta^* Ay + \theta^* x) - \lambda A\mathbf{tanh}(\beta^* x/\lambda + \theta^* x)\|. \quad (13)$$

We would like to bound each of the three terms on the right hand side by a constant times the left hand side of (11). For the first term, this is trivial. Further, we can show that the third term on the right hand side of (13) is bounded by the first term, using the Lipschitzness of tanh:

$$\|\lambda A\mathbf{tanh}(\beta^* Ay + \theta^* x) - \lambda A\mathbf{tanh}((\beta^*/\lambda)x + \theta^* x)\| \leq \|\lambda A\beta^* Ay - A\beta^* x\| \leq \|\beta^* A\|_2 \|x - \lambda Ay\| \\ \leq O(\|x - \lambda Ay\|),$$

where $\|\beta^* A\|_2 \leq O(1)$ using the assumptions of this paper. As for the second term, it represents the error of the mean field approximation for y, which corresponds to the substitution in (12). In order to bound this error term, we use the method of exchangeable pairs developed in [Cha05], which provides a strong and general concentration inequality for non-independent random variables. We can show that with high probability, this term will be $O(\|\lambda\beta^* A\|_2) \leq O(\lambda\|A\|_2) \leq O(\lambda\|A\|_F)$, since $B = O(1)$. Combining the above bounds we derive (11), as required.

3.2 Definitions of the terms in Theorem 6

We now sketch the proof of our general upper bound of Theorem 6. We first define the notions of covering numbers and the quantities $\mathcal{C}_1$ and $\mathcal{C}_2$ in the theorem statement.

Definition 1. *Given a metric space (Ω, d) and $\epsilon > 0$, a subset $\Omega' \subseteq \Omega$ is an* ϵ-net *for Ω if for any $\omega \in \Omega$ there exists $\omega' \in \omega$ such that $d(\omega, \omega') \leq \epsilon$. The covering number at scale ϵ, $N(\Omega, \epsilon)$, is the smallest size of an ϵ-net.*

For a function class $\mathcal{F}$ and collection of feature vectors $X = (x_1, \ldots, x_n)$, we denote by $N(\mathcal{F}, X, \epsilon)$ the covering number at scale ϵ of $\mathcal{F}$ w.r.t. the distance $d(f, g) = \sqrt{\|f(X) - g(X)\|_2^2/n}$, where we use the convenient notation $f(X) = (f(x_1), \ldots, f(x_n))$ and similarly for $g(X)$.

Next, we define the quantities $\mathcal{C}_1$ and $\mathcal{C}_2$. We start by defining the following as a function of $\beta, \beta' \in \mathbb{R}$ and $h, h' \in \mathbb{R}^n$:

$$\psi(h, \beta; h', \beta') = (\beta - \beta')^2 \|A\|_F^2 + \left\| h - h' + (\beta - \beta') A\mathbf{tanh}\left(\frac{\beta'}{\beta - \beta'}(h' - h) + h'\right)\right\|_2^2 \quad (14)$$

where $\mathbf{tanh}((z_1, \ldots, z_n)) = (\tanh(z_1), \ldots, \tanh(z_n))$. Now, for some universal constant $c \geq 0$, we define

$$\mathcal{C}_1(\mathcal{F}, X, \theta^*, \beta^*) := \sup_{(\theta,\beta)\in\Theta\times[-B,B]} \min\left(\frac{\|f_\theta(X) - f_{\theta^*}(X)\|_2^2/n}{\psi(f_\theta(X), \beta; f_{\theta^*}(X), \beta^*)}, \|f_\theta(X) - f_{\theta^*}(X)\|_2^2/n\right). \quad (15)$$

Similarly, $\mathcal{C}_2$ is defined in an analogous way, by replacing $\|f_\theta(X) - f_{\theta^*}(X)\|_2^2/n$ with $(\beta - \beta^*)^2$. Conveniently, we can use the following upper bound on $\mathcal{C}_1$:

$$\mathcal{C}_1'(\mathcal{F}, X, \theta^*, \beta^*) := \sup_{(\theta,\beta)\in\Theta\times[-B,B]} \frac{\|f_\theta(X) - f_{\theta^*}(X)\|_2^2/n}{\psi(f_\theta(X), \beta; f_{\theta^*}(X), \beta^*)}. \quad (16)$$

At this point, we can explain how the rate in (7) for $d = 1$ is derived from the bound of Theorem 6. In this case, $(x_1, \ldots, x_n)$ is simply a vector $x \in \mathbb{R}^n$ and $f_\theta(x_i) = \theta x_i$. Substituting $\mathcal{C}_1 \leq \mathcal{C}_1'$ into (5), substituting $f_\theta(x_i) = \theta x_i$ and substituting $\lambda = -(\beta - \beta^*)/(\theta - \theta^*)$, (7) follows.

3.3 Sketch of the upper bound in Theorem 6

Here, we sketch the proof of the upper bound in Theorem 6, but a weaker one where $\mathcal{C}_1$ is replaced by its upper bound $\mathcal{C}_1'$ defined in (15). In particular, we sketch that w.pr. $\geq 1-\delta$,

$$\frac{1}{n}\|f_{\hat{\theta}}(X) - f_{\theta^*}(X)\|_2^2 \lesssim \mathcal{C}_1'(\mathcal{F}, X, \beta^*, \theta^*) \inf_{\epsilon \geq 0} \left(\log \frac{n}{\delta} + \epsilon n + \log N(\mathcal{F}, X, \epsilon)\right). \tag{17}$$

It is possible to prove that $\mathcal{C}_1' \leq O(1/\|A\|_F^2)$, similarly to the corresponding argument in Section 3.1 and we focus below on proving (17).

Notice that in the definition of $\mathcal{C}_1$ and $\mathcal{C}_1'$, we do not need the set $\mathcal{F}$ itself, but only the vectors $f_\theta(X)$ for every θ in the class $\mathcal{F}$. Hence, if we define the set $\mathcal{H} = \{f_\theta(X) : f_\theta \in \mathcal{F}\}$, we immediately observe that $\mathcal{C}_1$ is in fact a function of $\mathcal{H}$. In this setting, we can similarly define $h^* = f_{\theta^*}(X)$ and $\hat{h} = f_{\hat{\theta}}(X)$ and define the covering numbers $N(\mathcal{H}, \epsilon)$ with respect to the distance $d(h, h') = \sqrt{\|h - h'\|_2^2/n}$. In this language, (17) translates to

$$\frac{1}{n}\|\hat{h} - h^*\|_2^2 \lesssim \mathcal{C}_1'(\mathcal{H}, h^*, \beta^*) \inf_{\epsilon \geq 0} \left(\log \frac{n}{\delta} + \epsilon n + \log N(\mathcal{H}, \epsilon)\right) \quad . \tag{18}$$

In the remainder of the proof, we will focus on proving (18), dividing the proof to multiple steps.

Step 1: A single dimensional $\mathcal{H}$. In this case, $\mathcal{H}$ is a single dimensional subspace of $\mathbb{R}^n$, namely, there exists $v \in \mathbb{R}^n$ such that $\mathcal{H} = \{h^* + tv \colon t \in \mathcal{T} \subseteq \mathbb{R}\}$. This is clearly reminiscent of the setting on a one-dimensional function-class discussed in Section 3.1. Hence, using the exact same approach and using the calculation of Section 3.2, we can prove that w.pr. $1-\delta'$

$$\frac{1}{n}\|\hat{h} - h^*\|_2^2 \lesssim \mathcal{C}_1'(\mathcal{H}, h^*, \beta^*) \log \frac{n}{\delta'}.$$

Step 2: A union of single-dimensional classes. Now, suppose that we have a finite set of directions (unit vectors) $v_1, \ldots, v_N$ and denote $\mathcal{H}_i = \{h^* + tv_i : h^* + tv_i \in \mathcal{H}\}$. In other words, $\mathcal{H}_i$ is the restriction of $\mathcal{H}$ on a specific line passing through h^* with direction v_i. Suppose we run MPLE on each direction, producing an output $\hat{h}_i$ for each direction. The calculations of Step 1 suggest that for all $i \in [N]$, w.pr. $1-\delta'$:

$$\frac{1}{n}\|\hat{h}_i - h^*\|_2^2 \lesssim \mathcal{C}_1'(\mathcal{H}, h^*, \beta^*) \log \frac{n}{\delta'}.$$

With a simple union bound over these N events, we can set $\delta' = \delta/N$ and obtain that w.pr. $\geq 1-\delta$, for all $i \in [N]$,

$$\frac{1}{n}\|\hat{h}_i - h^*\|_2^2 \lesssim \mathcal{C}_1'(\mathcal{H}, h^*, \beta^*) \left(\log \frac{n}{\delta} + \log N\right). \tag{19}$$

This essentially means that, if we run MPLE on the original set $\mathcal{H}$ and it ends up lying in any of the $\mathcal{H}_i$'s, it will lie close to the optimal point h^*.

Since we don't know in which direction the MPLE will lie, we have to establish a statement like (19) for all directions in $\mathcal{H}$. The problem is that usually there are infinity directions, so the union bound approach doesn't automatically work.

However, we can approximate the set of directions by a finite subset of directions that form an ϵ-net. Since any point $h \in \mathcal{H}$ defines a direction $h - h^*$, we can take an ϵ-net $\mathcal{U}$ with respect to $\mathcal{H}$, which has size $N = N(\mathcal{H}, \epsilon)$, which corresponds to the covering number defined in Definition 1 of Section 3.2. Due to Lipschitzness of the optimization target, one can prove that the MPLE over $\mathcal{U}$ is close to the MPLE over $\mathcal{H}$. By selecting ϵ appropriately and substituting $N = N(\mathcal{H}, \epsilon)$ in (19), we derive (18).

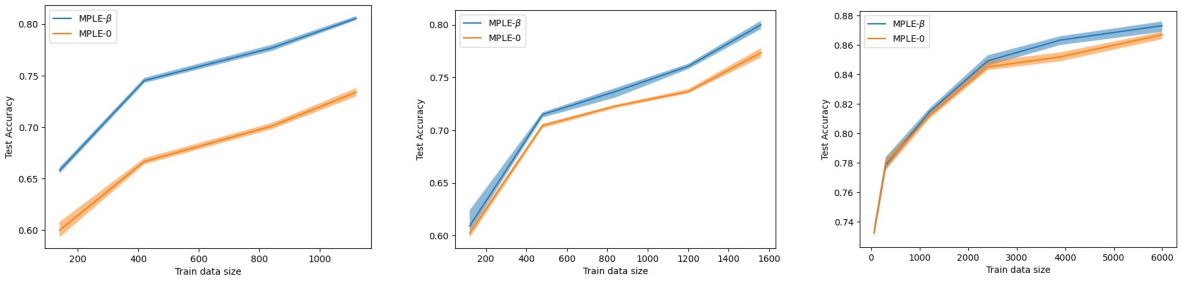

Figure 1: From Left to Right: Plots of the accuracy of MPLE-β (*blue*) vs MPLE-0 (*orange*) for Cora, Citeseer, Pubmed respectively as we increase the training data size gradually while maintaining the class probabilities.

Table 1: Datasets: Cora and Citeseer have probability vectors as features. Pubmed has TF-IDF frequencies as features.

Dataset	Classes	Nodes	Edges	Features
Cora	7	2708	5429	1433
Citeseer	6	3327	4732	3703
Pubmed	3	19717	44338	500

4 Experiments

When there is network information about dependencies between samples, we can use it to significantly boost the performance of supervised learning approaches. We demonstrate such improvements of MPLE-β from including the dependency structure compared to assuming the data is i.i.d. We call this MPLE-0 (i.e. setting $\beta = 0$). We observe that MPLE-β consistently outperforms MPLE-0 by a significant margin.

Datasets. We utilize three public citation datasets - Cora, Citeseer and Pubmed [YCS16]. These datasets consist of a network where each node corresponds to a publication and the edges correspond to citation links. Each node contains a bag-of-words representation of the publication and a corresponding label indicating the area of the publication. Table 1 gives the statistics of the datasets.

Experimental Setup. The datasets we use are common benchmarks used for semi-supervised and fully-supervised learning on graph structured data. The state of the art on a lot of these is graph neural network (GNN) [Che+20] based approaches. The setups considered in prior literature on these datasets differ from ours in the following sense: these works consider the *transductive* setting, that is, they assume access to the adjacency matrix of the entire graph as well as the features of the entire dataset (including those in the test set) at train time. In contrast, we work in the *inductive* setting, where we do not assume access to any information about the test set. However, at test time, our hypothesis uses the labels in the validation set (not the features).

We perform three different experiments on each dataset where we measure the accuracy of prediction on the test labels. We run each experiment with 10 fixed random seeds and report the average and standard deviation.

1. *Sparse-data*: Following the semi-supervised setup of [KW16; Fen+20] and others, we compare performance of MPLE-0 and MPLE-β over a public split which includes only 20 nodes per class as training, 500 nodes for validation and 1000 nodes for testing.

2. *Increasing training data*: We compare the gap in performance of the two methods when training data is gradually increased from the semi-supervised setting towards the full-supervised setting.

3. *Full-supervised*: We consider the fully-supervised setup from [Pei+20]. In this setup, we consider 10 random splits of the entire dataset. Each split maintains class distribution by splitting the set of nodes of

Table 2: Accuracy comparison between MPLE-0, MLPE-β and GCNII-In for full-supervised experiment.

Dataset	MPLE-0	MPLE-β	GCNII-In
Cora	74.5 ± 1.8	$\mathbf{85.3} \pm 1.7$	$\mathbf{85.3} \pm 1.3$
Citeseer	72.3 ± 1.7	$\mathbf{76.3} \pm 1.0$	68.6 ± 0.3
Pubmed	87.3 ± 0.2	$\mathbf{89.0} \pm 0.2$	83.3 ± 0.6

each class into 60%(train)-20%(val)-20%(test). For this experiment, we compare against an inductive variant of GCNII we denote GCNII-In. We disable access to the test set features during training in order to have a fair comparison with our inductive setting.

Model Details. Since our classification task is multi-class, we extend the MPLE-β algorithm for Ising models to its natural Pott's model generalization. For number of classes K, the probability of label $y_i = k^*$ conditioned on the other data and labels is computed as follows:

$$\mathbb{P}_{\theta,\beta}[y_i = k^* | x, y_{-i}] = \frac{\exp\left(f_\theta(x_i)_{k^*} + \beta \sum_{j=1}^n A_{ij} \mathbb{1}[y_j = k^*]\right)}{\sum_{k=1}^K \exp\left(f_\theta(x_i)_k + \beta \sum_{j=1}^n A_{ij} \mathbb{1}[y_j = k]\right)}.$$

Using this we compute the MPLE-β objective.

For both MPLE-0 and MPLE-β, our underlying model $f_\theta : \mathbb{R}^{\#\mathsf{features}} \to \mathbb{R}^{\#\mathsf{classes}}$ is a 2-layer neural network with 32 units in the hidden layer and ReLU activations. The difference between the two models is just the use of β. For comparison with the graph neural networks (GNNs), we use the GCNII [Che+20] model which is a state-of-the-art GNN with depth 64 and hidden layer size of 64. We run our code on a GPU and use Adam to train all our models. We use the tuned hyper-parameters for GCNII however for our algorithms we do not perform a hyper-parameter search but use the parameters used in prior work [Fen+20].

Results. On the sparse-data experiment, for Cora MPLE-β gives an accuracy of $65.8 \pm 0.09\%$ vs $60 \pm 0.4\%$ given by MPLE-0. For Citeseer, MPLE-β gets $60.9 \pm 0.7\%$ vs MPLE-0 which gets $60.2 \pm 0.3\%$. For Pubmed, both approaches get $73.3 \pm 0.2\%$. As we increase the train data size as shown in Figure 1 our gains also tend to increase. Finally for the fully-supervised setting we again outperform MPLE-0 *and* GCNII-In. On Pubmed, our gains are smaller as the TF-IDF feature vector already implicitly encodes some network information from the neighbors. Moreover, MPLE-β runs much faster than any of the GNN approaches and is simpler with a low overhead of a scalar parameter on any given model, while remaining competitive in performance. However, it should be noted that we do not compare performance in the transductive setting, in which GCNII was probably intended to run. Finally, our experiments are based on an approach with provable end-to-end guarantees, in contrast with the GNN approaches.

A Preliminary: formulation as an Ising model

We start by repeating some of the definitions from Section 3.2, and proceed by presenting a modified notation, that we will use in the proof of Theorem 6, as it is easier to handle.

Central definitions from Section 3.2.

Definition 2. *Let $h, h' \in \mathbb{R}^n$ and $\beta, \beta' \in \mathbb{R}$. We define*

$$\psi(h, \beta; h', \beta') = (\beta - \beta')^2 \|A\|_F^2 + \left\| h - h' + (\beta - \beta') A \mathbf{tanh}\left(\frac{\beta'}{\beta - \beta'}(h' - h) + h'\right)\right\|_2^2$$

Next, we define the quantities $\mathcal{C}_1, \mathcal{C}_2$ that appear in the rate of Theorem 6.

Definition 3. *For a function class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ and a collection of feature vectors $X \in \mathbb{R}^{n\times d}$, where the i-th row of X is the feature x_i, we define*

$$\mathcal{C}_1(\mathcal{F}, X, \theta^*, \beta^*) := \sup_{(\theta,\beta)\in\Theta\times[-B,B]} \min\left(\frac{\|f_\theta(X) - f_{\theta^*}(X)\|_2^2/n}{\psi(f_\theta(X), \beta; f_{\theta^*}(X), \beta^*)}, \|f_\theta(X) - f_{\theta^*}(X)\|_2^2/n\right)$$
$$\mathcal{C}_2(\mathcal{F}, X, \theta^*, \beta^*) := \sup_{(\theta,\beta)\in\Theta\times[-B,B]} \min\left(\frac{(\beta-\beta^*)^2}{\psi(f_\theta(X), \beta; f_{\theta^*}(X), \beta^*)}, (\beta-\beta^*)^2\right)$$

In some parts of the proof, it will be more convenient to work with the following simplified upper bounds to $\mathcal{C}_1$ and $\mathcal{C}_2$:

$$\mathcal{C}_1'(\mathcal{F}, X, \theta^*, \beta^*) := \sup_{(\theta,\beta)\in\Theta\times[-B,B]} \frac{\|f_\theta(X) - f_{\theta^*}(X)\|_2^2/n}{\psi(f_\theta(X), \beta; f_{\theta^*}(X), \beta^*)}$$
$$\mathcal{C}_2'(\mathcal{F}, X, \theta^*, \beta^*) := \sup_{(\theta,\beta)\in\Theta\times[-B,B]} \frac{(\beta-\beta^*)^2}{\psi(f_\theta(X), \beta; f_{\theta^*}(X), \beta^*)}$$

Definition 4. *Given $\epsilon > 0$, $\mathcal{F}$ and X, denote by $\mathcal{N}(\epsilon, \mathcal{F})$ the ϵ-covering number of $\mathcal{F}$ with respect to the distance*

$$d(f, f') = ||f(X) - f'(X)||/\sqrt{n} = \sqrt{\frac{1}{n}\sum_{i=1}^n (f(x_i) - f'(x_i))^2}.$$

In other words, $\mathcal{N}(\epsilon, \mathcal{F})$ is the minimal cardinality of a set $\mathcal{G} \subseteq \mathcal{F}$, such that for any $f \in \mathcal{F}$ there exists $g \in \mathcal{G}$ such that $d(f, f') \le \epsilon$.

A modified notation for the proof. To prove Theorem 6 we use a slightly modified setting, replacing the family of vectors $\{f_\theta(X) \colon \theta \in \Theta\}$ with a family $\mathcal{H}$ of elements of $\mathbb{R}^n$. This comes from the realization that the only way a function f_θ influences the outcome is thought the vector $(f_\theta(x_1), \dots, f_\theta(x_n)) \in \mathbb{R}^n$. Hence, it is more convenient to consider the set of all these vectors that are produced for various θ as our main object of interest. We have a fixed matrix A with $\|A\|_\infty = 1$. We also replace $(y_1, \dots, y_n)$ by a vector $\sigma \in \{-1, 1\}^n$ and for any $h \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$, denote

$$\mathbb{P}_{h,\beta}[\sigma] = \frac{1}{Z_{h,\beta}} \exp\left(\beta\sigma^{\mathrm{T}} A\sigma + \sum_{i=1}^n \sigma_i h_i\right).$$

We can also write the MPLE Eq. (5) in this language. Define the negative log of the optimized quantity by

$$\varphi(h, \beta; \sigma) := -\log \prod_{i=1}^n \mathbb{P}_{h,\beta}[\sigma_i | \sigma_{-i}]$$

and by $(\hat{h}, \hat{\beta})$ the MPLE, namely,

$$(\hat{h}, \hat{\beta}) = \arg\min_{h\in\mathcal{H}, \beta\in[-B,B]} \varphi(h, \beta; \sigma).$$

We define $\mathcal{C}_1$ and $\mathcal{C}_2$ analogously with respect to $\mathcal{H}$ instead of $\mathcal{F}$:

Definition 5. *Suppose $\mathcal{H} \subseteq \mathbb{R}^n$ and let $\beta^* \in \mathbb{R}, h^* \in \mathbb{R}^n$ be fixed. We define*

$$\mathcal{C}_1(\mathcal{H}, h^*, \beta^*) := \sup_{(h,\beta)\in\mathcal{H}\times[-B,B]} \min\left(\frac{\|h - h^*\|_2^2/n}{\psi(h, \beta; h^*, \beta^*)}, \|h - h^*\|_2^2/n\right)$$
$$\mathcal{C}_2(\mathcal{H}, h^*, \beta^*) := \sup_{(h,\beta)\in\mathcal{H}\times[-B,B]} \min\left(\frac{(\beta-\beta^*)^2}{\psi(h, \beta; h^*, \beta^*)}, (\beta-\beta^*)^2\right)$$

Similarly, we define $\mathcal{N}(\epsilon, \mathcal{H})$:

Definition 6. *Given $\mathcal{H}$ and $\epsilon > 0$, denote by $\mathcal{N}(\epsilon, \mathcal{H})$ the ϵ-covering number of $\mathcal{H}$ with respect to the distance $d(h, h') = \|h - h'\|/\sqrt{n}$.*

In various parts of the proof, we will use symbols like c, C, C'. These denote constants that possibly depend exponentially in M and B and are independent of the other parameters.

Re-writing the main theorem in the modified notation. We can rewrite Theorem 6 in the modified notation. We split it to Theorem 7 and Theorem 8, proving the upper and lower bounds, respectively.

The following theorem is proved in Section B:

Theorem 7. *Let A be a matrix of $\|A\|_\infty = 1$, let $M, B > 0$, let $\mathcal{H} \subseteq [-M, M]^n$ and let $\beta^* \in [-B, B]$. Let $(\hat{h}, \hat{\beta})$ denote the MPLE over $\mathcal{H} \times [-M, M]$. Then, there is a constant $C(M, B)$ that depends singly exponentially on M, B, such that for any $\delta \in (0, 1/2)$, with probability at least $1 - \delta$, we have*

$$\frac{\|\hat{h} - h^*\|^2}{n} \leq C(M, B) \inf_{\epsilon \geq 0} \left(\log \frac{n}{\delta} + \epsilon n + \log \mathcal{N}(\epsilon, \mathcal{H}) \right) \mathcal{C}_1(\mathcal{H}, h^*, \beta^*)$$

and

$$(\hat{\beta} - \beta^*)^2 \leq C(M, B) \inf_{\epsilon \geq 0} \left(\log \frac{n}{\delta} + \epsilon n + \log \mathcal{N}(\epsilon, \mathcal{H}) \right) \mathcal{C}_2(\mathcal{H}, h^*, \beta^*).$$

The following theorem is proved in Section C:

Theorem 8. *Let A be a matrix of $\|A\|_\infty = 1$, let $M, B > 0$, let $\mathcal{H} \subseteq [-M, M]^n$ and let $\beta^* \in [-B, B]$. Assume that $\mathcal{H}$ is convex and closed under negation, namely, for any $h, h' \in \mathcal{H}$ and $\lambda \in (0, 1)$ it holds that $\lambda h + (1 - \lambda) h' \in \mathcal{H}$ and $-h \in \mathcal{H}$. Then, for any estimator h' there exists $(h^*, \beta^*) \in \mathcal{H} \times [-B, B]$ s.t. with probability ≥ 0.49 over $\sigma \sim \mathbb{P}_{h^*, \beta^*}$,*

$$\frac{\|h' - h^*\|^2}{n} \geq c(M) \mathcal{C}_1(\mathcal{H}, h^*, \beta^*)$$

where $C(M)$ depends exponentially on M.

We also have the following simple Lemma that upper bounds $\mathcal{C}_1, \mathcal{C}_2$ by a simpler quantity, and proved in Section B.4.

Lemma 1. *It holds that*

$$\mathcal{C}_1(\mathcal{H}, h^*, \beta^*) \leq \frac{1}{4\|A\|_F^2}.$$

Proof of Theorem 6. We now explain how we can easily use Theorem 7, Theorem 8 and Lemma 1 to get Theorem 6.

Proof of Theorem 6. For the upper bound, we will apply Theorem 6 by setting $\mathcal{H} = \{(f_\theta(x_1), \ldots, f_\theta(x_n)) : \|\theta\| \leq M\}$. First of all, we show that $N(\mathcal{F}, X, \epsilon) = N(\mathcal{H}, \epsilon)$. Indeed, by the definition of $\mathcal{H}$ we can see how the elements of $\mathcal{F}$ can be in one-to-one correspondence with the elements of $\mathcal{F}$. Specifically, some $f_\theta \in \mathcal{F}$ corresponds to $h = (f_\theta(x_1), \ldots, f_\theta(x_n)) \in \mathcal{H}$. As for the metric, if h_1, h_2 correspond to $f_{\theta_1}, f_{\theta_1}$, we have

$$d(f_{\theta_1}, f_{\theta_2}) = \sqrt{\sum_{i=1}^n (f_{\theta_2}(x_i) - f_{\theta_1}(x_i))^2 / n} = \frac{\|h_2 - h_1\|}{\sqrt{n}}$$

Thus, we see that the metric d corresponds exactly to the metric used to define $N(\mathcal{H}, \epsilon)$ in Section 3.3. It follows that $N(\mathcal{F}, X, \epsilon) = N(\mathcal{H}, \epsilon)$.

Since optimizing in the space $\mathcal{H}$ is equivalent to optimizing in the space of θ's, we have $\hat{h} = f_{\hat{\theta}}(X)$ and $h^* = f_{\theta^*}(X)$. To conclude the proof of the general rate, we need to argue that $\mathcal{C}_1(\mathcal{H}, h^*, \beta^*) = \mathcal{C}_1(\mathcal{F}, X, \theta^*, \beta^*)$. Since in the definition of $\mathcal{C}_1(\mathcal{H}, h^*, \beta^*)$ we just substituted $h = f_\theta(X)$ and to create the set $\mathcal{H}$ we did the same thing, these two quantities are clearly the same. A similar argument can be made for $\mathcal{C}_2(\mathcal{H}, h^*, \beta^*)$.

To conclude the proof of the upper bound in Theorem 6, we simply use Lemma 1 to bound the quantities $\mathcal{C}_1, \mathcal{C}_2$. The lower bound similarly follows from Theorem 8. □

B Proof of the upper bound in Theorem 6

B.1 Proof Overview

As was shown in Section A, it suffices to prove Theorem 7. The proof can be broken down into several lemmas. We first present a lemma that shows that for a single $h \in \mathcal{H}$, if h is far from h^* it is unlikely that the MPLE will return h.

Lemma 2. *Let $h \in [-M, M]^n$ be a fixed vector and $\beta \in [-B, B]$. Then, there is a constant $C = C(M, B) > 0$ such that for all $u > 0$, with probability at least $1 - \log n \mathrm{e}^{-u^2}$, if*

$$\|(\beta - \beta^*)A\sigma + h - h^*\|_2 \geq Cu \tag{20}$$

then

$$\varphi(h, \beta; \sigma) > \varphi(h^*, \beta^*; \sigma) + cu^2 \tag{21}$$

Proof. First, denote

$$\mathbf{a} = (h - h^*, \beta - \beta^*)$$

where $\mathbf{a} \in \mathbb{R}^{n+1}$. Define the function $g : [0, 1] \to \mathbb{R}$ as

$$g(t) = \varphi(h^* + t(h - h^*), \beta^* + t(\beta - \beta^*); \sigma) = \varphi((h^*, \beta^*) + t\mathbf{a}; \sigma)$$

For simplicity, from now on we suppress the dependence of φ on the sample σ since it is obvious.

Consider the Taylor expansion of g around 0. Notice also that $g(0) = \varphi(h^*, \beta^*)$. We have

$$\begin{aligned} g(t) &= g(0) + tg'(0) + \frac{t^2}{2}g''(t') \\ &= \varphi(h^*, \beta^*) + t\mathbf{a}^\top \nabla\varphi(h^*, \beta^*) + \frac{t^2}{2}\mathbf{a}^\top \nabla^2\varphi(h', \beta')\mathbf{a}, \end{aligned}$$

where $t' \in [0, t]$ and $(h', \beta') = (h^*, \beta^*) + t'\mathbf{a}$. It becomes clear that in order to control the quantity on the right, we need to understand the relation of the first and the second order terms of the Taylor. To do that, we extend some concentration bounds from [Dag+20] involving the first and second derivatives, as follows (proof in Section B.3):

Lemma 3. *Let $t \geq 0$, $\beta \in \mathbb{R}$ and $h \in \mathbb{R}^n$. Then, with probability at least $1 - \log n \exp(-ct^2)$, if*

$$\|\beta A\sigma + h\|_2 \geq Ct|\beta|\|A\|_2$$

then

$$\frac{\left|(h, \beta)^\top \nabla\varphi(h^*, \beta^*)\right|}{\min_{\beta' \in [-M,M], h' \in [-M,M]^n} (h, \beta)^\top \nabla^2\varphi(h', \beta')(h, \beta)} \leq \frac{c't}{\|\beta A\sigma + h\|_2}$$

Using the result of this lemma, we get that with probability $1 - \log n \mathrm{e}^{-u^2}$, if

$$\|(\beta - \beta^*)A\sigma + h - h^*\|_2 \geq Cu\|A\|_2|\beta - \beta^*|, \tag{22}$$

then

$$\frac{|\mathbf{a}^\top \nabla \varphi(J^*, h^*)|}{\mathbf{a}^\top \nabla^2 \varphi(h', \beta')\mathbf{a}} \leq \frac{Cu}{\|(\beta - \beta^*)A\sigma + h - h^*\|_2}. \tag{23}$$

Notice that (20) implies (22) (if the constant $C > 0$ in (22) is sufficiently large). Then, with probability $1 - \log n e^{-u^2}$, (20) implies (23). Hence, it suffices to prove that (23) implies (21). For the remainder of the proof, assume that (23) holds. Substituting back to the Taylor expression, we get that

$$\begin{aligned}\varphi(h, \beta) - \varphi(h^*, \beta^*) = g(1) + g(0) &\geq \frac{1}{2}\mathbf{a}^\top \nabla^2 \varphi(h', \beta')\mathbf{a} - \left|\mathbf{a}^\top \nabla \varphi(h^*, \beta^*)\right| \\ &\geq \mathbf{a}^\top \nabla^2 \varphi(h', \beta')\mathbf{a}\left(\frac{1}{2} - \frac{C'u}{\|(\beta - \beta^*)A\sigma + h - h^*\|}\right) \geq \frac{1}{4}\mathbf{a}^\top \nabla^2 \varphi(h', \beta')\mathbf{a},\end{aligned}$$

where the last inequality holds if the constant C in (20) is sufficiently large. To bound the right hand side, we use the following lemma that will be proven in Section B.3 using the techniques of [Dag+20]:

Lemma 4. *Let $\beta' \in [-B, B]$, $h' \in [-M, M]^n$, $\beta \in \mathbb{R}$ and $h \in \mathbb{R}^n$. Then,*

$$(h, \beta)^\top \nabla^2 \varphi(h', \beta')(h, \beta) \geq c\|\beta A\sigma + h\|_2.$$

As a consequence of Lemma 4 and (20),

$$\mathbf{a}^\top \nabla^2 \varphi(h', \beta')\mathbf{a} \geq c\|(\beta - \beta^*)A\sigma + h - h^*\|_2^2 \geq cu^2,$$

which proves (21) and concludes the proof. □

After establishing Lemma 2, it is useful to examine what guarantees we get from it. If we take the contrapositive of the statement, it means that if the MPLE algorithm returns the vector h as an estimate, then with high probability Eq. (20) holds. However, this inequality is a random quantity that depends on the particular instance of σ that we sample. Thus, we want to replace it with a nonrandom inequality, which is the content of the next lemma. We use the quantity $\psi(h, \beta; h', \beta')$ defined in (20) and write use the shorthand notation

$$\psi(h, \beta) := \psi(h, \beta; h^*, \beta^*).$$

Lemma 5. *Let $h \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$. With probability $1 - C \log n e^{-c\psi(h,\beta)/(\beta - \beta^*)\|A\|_2^2}$,*

$$\|h^* - h - (\beta - \beta^*)A\sigma\|_2^2 \geq c\psi(h, \beta).$$

The proof of Lemma 5 is divided into two parts. Since $\psi(h, \beta)$ involves two terms, each part consists of showing that the normed quantity is greater than each of the two terms. For the first term, we have the following Lemma.

Lemma 6. *Let $h \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$. Then, with probability at least $1 - \log n \exp(-c\|A\|_F^2/\|A\|_2^2)$*

$$\|h^* - h + (\beta^* - \beta)A\sigma\|_2 \geq c|\beta - \beta^*|\|A\|_F.$$

This is proved in Section B.3, and follows from arguments from [Dag+20]. The idea is to use concentration bounds about second degree polynomials of an Ising model. These immediately yield the required dependence on $\|A\|_F$.

For the second part of the proof of Lemma 5, we need the following Lemma.

Lemma 7. *Let $h \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$. Then, with probability at least*

$$1 - \exp\left(-c\frac{\left\|h^* - h + (\beta^* - \beta)A\tanh\left(\frac{\beta^*}{\beta - \beta^*}(h^* - h) + h^*\right)\right\|_2^2}{(\beta - \beta^*)^2\|A\|_2^2}\right)$$

we have that

$$\|h^* - h + (\beta^* - \beta)A\sigma\|_2 \geq c \left\| h^* - h + (\beta^* - \beta)A \tanh\left(\frac{\beta^*}{\beta - \beta^*}(h^* - h) + h^*\right)\right\|_2 .$$

Before proving Lemma 7, we need some auxiliary lemma about the concentration of a function of the Ising model. This lemma is proved in Section B.2 using the technique of exchangeable pairs [Cha05].

Lemma 8. *Let $\sigma \sim \mathbb{P}_{h^*,\beta^*}$, let $v \in \mathbb{R}^n$ and let*

$$f(\sigma) = \sum_{i=1}^{n} v_i(\sigma_i - \tanh(\beta^*(A\sigma)_i + h_i^*).$$

Then, for all $t \geq 0$,

$$\mathbb{P}[|f(\sigma)| > t] \leq 2\exp\left(-\frac{t^2}{8\|v\|_2^2\,(1 + |\beta^*|\|A\|_\infty)}\right).$$

Proof of Lemma 7. Denote $\tilde{\beta} = \beta - \beta^*$. First of all, notice that we can rewrite the quantity of interest as

$$\|h^* - h - \tilde{\beta}A\sigma\|$$

The strategy for bounding this quantity will be to use the mean field approximation for the Ising model. Of course, we have to show that this is a good enough approximation in our case. Let $\mathbf{u} \in \mathbb{R}^n$ be an arbitrary vector with $\|\mathbf{u}\| = 1$. We start by using the triangle inequality to write

$$\left|\mathbf{u}^\top\left(h^* - h - \tilde{\beta}A\tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^* - h) + h^*\right)\right)\right| \leq |\mathbf{u}^\top(h^* - h - \tilde{\beta}A\sigma)| + |\mathbf{u}^\top\left(\tilde{\beta}A\sigma - \tilde{\beta}A\tanh\left(\beta^*A\sigma + h^*\right)\right)| +$$
$$+ \left|\mathbf{u}^\top\left(\tilde{\beta}A\tanh\left(\beta^*A\sigma + h^*\right) - \tilde{\beta}A\tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^* - h) + h^*\right)\right)\right|$$

We will bound each of these three terms separately. For the first term, using the Cauchy-Schwarz inequality, we obtain

$$|\mathbf{u}^\top(h^* - h - \tilde{\beta}A\sigma)| \leq \|\mathbf{u}\|\|h^* - h - \tilde{\beta}A\sigma\| = \|h^* - h - \tilde{\beta}A\sigma\| \tag{24}$$

We proceed with bounding the second term. Define $\mathbf{v} = \tilde{\beta}A^\top\mathbf{u}$. Then this term can be written as

$$|\mathbf{v}^\top(\sigma - \tanh(\beta^*A\sigma + h^*))| = \left|\sum_{i=1}^{n} v_i(\sigma_i - \tanh(\beta^*(A\sigma)_i + h_i^*))\right|.$$

We use Lemma 8 to derive that for any $t \geq 0$, this term can be bounded by $Ct\|v\|_2\sqrt{1 + |\beta^*|\|A\|_\infty}$ with probability e^{-ct^2}. Using the facts that $|\beta^*|, \|A\|_\infty \leq O(1)$ and

$$\|v\|_2 = \|\tilde{\beta}u^\top A\|_2 \leq |\tilde{\beta}|\|u\|_2\|A\|_2 = |\tilde{\beta}|\|A\|_2$$

to derive that w.p. $1 - e^{-t^2}$,

$$|\mathbf{v}^\top(\sigma - \tanh(\beta^*A\sigma + h^*))| \leq Ct|\tilde{\beta}|\|A\|_2.$$

For the third term, we use Cauchy Schwarz inequality and the lipschitzness of tanh to conclude that

$$\left|\mathbf{u}^\top\left(\tilde{\beta}A\tanh\left(\beta^*A\sigma + h^*\right) - \tilde{\beta}A\tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^* - h) + h^*\right)\right)\right| \leq \beta^*\|A\|_\infty\|\mathbf{u}\|\|h^* - h - \tilde{\beta}A\sigma\|$$
$$\leq C\|h^* - h - \tilde{\beta}A\sigma\|,$$

using the fact that $\|u\|_2 = 1$ and that $|\beta^*|, \|A\|_\infty \leq 1$. Putting everything together, we conclude that for every $\mathbf{u} \in \mathbb{S}^{n-1}$, with probability at least $1 - \mathrm{e}^{-t^2}$

$$\left|\mathbf{u}^\top \left(h^* - h - \tilde{\beta} A \tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^* - h) + h^*\right)\right)\right| \leq C' t|\tilde{\beta}|\|A\|_2 + C'\|h^* - h - \tilde{\beta} A\sigma\|$$

Setting

$$\mathbf{u} = \frac{h^* - h - \tilde{\beta} A \tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^* - h) + h^*\right)}{\left\|h^* - h - \tilde{\beta} A \tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^* - h) + h^*\right)\right\|},$$

which is a deterministic vector, we get that with probability $1 - \mathrm{e}^{-t^2}$

$$\left\|h^* - h - \tilde{\beta} A \tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^* - h) + h^*\right)\right\| \leq C\|h^* - h - \tilde{\beta} A\sigma\| + Ct|\tilde{\beta}|\|A\|_2 \implies$$
$$\|h^* - h - \tilde{\beta} A\sigma\| \geq c_1 \left\|h^* - h - \tilde{\beta} A \tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^* - h) + h^*\right)\right\| - C_2 t|\tilde{\beta}|\|A\|_2$$

where c_1, C_2 are constants. By substituting $t = c\left\|h^* - h - \tilde{\beta} A \tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^* - h) + h^*\right)\right\| / |\tilde{\beta}|\|A\|_2$ for a sufficiently small constant $c > 0$, the proof follows. □

Given Lemmas 6 and 7, we can finally prove Lemma 5.

Proof of Lemma 5. We divide into cases: if

$$|\beta - \beta^*|\|A\|_F \geq \left\|h^* - h - (\beta - \beta^*) A \tanh\left(\frac{\beta^*}{\beta - \beta^*}(h^* - h) + h^*\right)\right\|_2$$

then the proof follows directly from Lemma 6, otherwise it follows directly from Lemma 7. □

Using Lemma 5 we can show that if $\psi(h, \beta)$ is large, then the pseudo-likelihood value for (h, β) will be far from optimal. This is the first indication that maximizing the likelihood might give us meaningful information about the parameters.

Lemma 9. *Let $h \in [-M, M]^n$ and $\beta \in [-B, B]$. Then, with probability $1 - \log n\, \mathrm{e}^{-c\psi(h,\beta)}$,*

$$\varphi(h, \beta) \geq \varphi(h^*, \beta^*) + c\psi(h, \beta).$$

It follows that for any $u > 0$ with $u^2 \leq \psi(h, \beta)$ we have with probability at least $1 - \log n\, \mathrm{e}^{-cu^2}$

$$\varphi(h, \beta) \geq \varphi(h^*, \beta^*) + cu^2.$$

Proof. First of all, we have from Lemma 5 that with probability at least

$$1 - C \log n\, \mathrm{e}^{-c\psi(h,\beta)/(\beta-\beta^*)^2\|A\|_2^2} \geq 1 - C \log n\, \mathrm{e}^{-c'\psi(h,\beta)},$$

it holds that

$$\|h^* - h - (\beta - \beta^*) A\sigma\|_2 \geq c\psi(h, \beta). \tag{25}$$

Further, from Lemma 2, with probability $1 - \log n \exp(-c''\psi(h, \beta))$, if (25) holds then $\varphi(h, \beta) \geq \varphi(h^*, \beta^*) + c''\psi(h, \beta)$. This concludes the proof of the first claim. For the second claim, since $u^2 \leq \psi(h, \beta)$, we have $1 - \log n \exp(-c\psi(h, \beta)) \geq 1 - \log n \exp(-cu^2)$, while at the same time

$$\varphi(h, \beta) \geq \varphi(h^*, \beta^*) + c\psi(h, \beta) \implies \varphi(h, \beta) \geq \varphi(h^*, \beta^*) + cu^2.$$

□

So far, we have focused on a specific (h, β) in the space of possible solutions to the MPLE problem. Next, we would like to show that the MPLE will satisfy this inequality. In order to do so, we would like to prove a high probability statement for all such pairs from $\mathcal{H} \times [-B, B]$. To prove such a statement, we need to make use of the properties of the metric space induced by $\mathcal{H}$ and $\|\cdot\|$, as stated in the following lemma:

Lemma 10. *With probability* $1 - \delta$,

$$\psi(\hat{h}, \hat{\beta}) \leq C \inf_{\epsilon \geq 0} \left(\epsilon n + \log \frac{n}{\delta} + \log \mathcal{N}(\epsilon, \mathcal{H}) \right). \tag{26}$$

Proof. Lemma 9 states that for a single pair (h, β), if $\psi(h, \beta)$ is large then with high probability, $\varphi(h, \beta) > \varphi(h^*, \beta^*)$. We need to argue that this holds for all pairs (h, β) such that $\psi(h, \beta)$ is larger than the right hand side of (26) and this will imply that the MPLE satisfies that $\psi(\hat{h}, \hat{\beta})$ is smaller that that quantity, as required. For convenience, let u denote the right hand side of (26).

A simple way to do this would be to use a union bound over all the points such that $\psi(h, \beta) \geq u$. Unfortunately, the set of these points is potentially uncountably infinite, hence this approach is infeasible. A common tool to bypass this obstacle is to define an ϵ-net over this set. If this net has a finite number of points, we can take the union bound over these points. This will mean that all the points in the net have a large value of φ. If the radius ϵ is chosen sufficiently small and the function φ is Lipschitz, then this implies that for all the points that are far, their value of φ is large.

We now start to execute on that strategy. First of all, we need to know the size of the ϵ-net of the set of points $\mathcal{H}_u := \{(h, \beta) \in \mathcal{H} \times [-M, M] \colon \psi(h, \beta) \geq u^2\}$. We can bound this in terms of the covering numbers for $\mathcal{H} \times [-M, M]$ using the following lemma, easily proven using the definition of covering numbers.

Lemma 11. *Let* (U, d) *be a metric space with domain* U *and metric function* d *and let* $V \subseteq U$*. Then, for any* $\epsilon > 0$

$$\mathcal{N}(\epsilon, V, d) \leq \mathcal{N}(\epsilon/2, U, d).$$

For reasons that will become apparent shortly, for $\mathcal{H}$ we will use the distance

$$d(h_1, h_2) = \frac{\|h_2 - h_1\|_2}{\sqrt{n}},$$

which was defined in Section 3.3. Hence, the metric for the whole space $\mathcal{H} \times [-B, B]$ will be

$$d_1((h_1, \beta_1), (h_2, \beta_2)) := d(h_1, h_2) + |\beta_2 - \beta_1|$$

The reason we define this metric is explained by the difference in the Lipschitz constants of φ with respect to the two parameters. Specifically, we have the following Lemma.

Lemma 12. *Let* $\varphi : \mathcal{H} \times [-M, M] \to \mathbb{R}$ *be the negative log pseudo-likelihood function. Then, for a fixed* h, φ *is* $2\|A\|n$*-Lipschitz with respect to* β *and for a fixed* β, φ *is* $2\sqrt{n}$*-Lipschitz with respect to* h *in* l_2*-norm. As a result,* φ *is* $2\|A\|n$*-Lipschitz with respect to both* h, β *and distance* d_1*.*

Proof. For a fixed β, let $h_1, h_2 \in \mathcal{H}$. Define the function $g : [0, 1] \to \mathbb{R}$ as

$$g(t) = \varphi(t h_2 + (1 - t) h_1, \beta)$$

and denote $h(t) = t h_2 + (1 - t) h_1$. we have that

$$\begin{aligned} |g'(t)| &= \left| \sum_{i=1}^{n} ((h_2)_i - (h_1)_i)(\sigma_i - \tanh(\beta A_i \sigma + h(t)_i)) \right| \\ &\leq \|h_2 - h_1\| \sqrt{\sum_{i=1}^{n} (\sigma_i - \tanh(\beta A_i \sigma + h(t)_i))^2} \end{aligned}$$

$$\leq 2\sqrt{n}\|h_2 - h_1\| \leq 2n\|h_2 - h_1\|_\infty,$$

where we have used the Cauchy-Schwarz inequality in the above calculations. Notice that this bound holds for all $t \in [0,1]$. Hence, by the mean value Theorem, we have

$$|\varphi(h_2,\beta) - \varphi(h_1,\beta)| = |g(1) - g(0)| = |g'(\xi)| \leq 2\sqrt{n}\|h_2 - h_1\|$$

for some $\xi \in (0,1)$. This shows that φ is $2\sqrt{n}$ Lipschitz for a fixed β. Now, suppose h is fixed. We define $r : [0,1] \to \mathbb{R}$ as

$$r(t) = \varphi(h, \beta(t))$$

where $\beta(t) = t\beta_2 + (1-t)\beta_1$. Similar to the previous computation, we have

$$\begin{aligned} r'(t) &= \left| (\beta_2 - \beta_1) \sum_{i=1}^{n} A_i \sigma(\sigma_i - \tanh(\beta(t) A_i \sigma + h_i)) \right| \\ &\leq |\beta_2 - \beta_1| \|A\sigma\| 2\sqrt{n} \\ &\leq 2\|A\| n |\beta_2 - \beta_1| \end{aligned}$$

where we again used Cauchy-Schwarz. Finally, by the mean value Theorem

$$|\varphi(h,\beta_2) - \varphi(h,\beta_1)| = |r(1) - r(0)| \leq 2\|A\| n |\beta_2 - \beta_1|$$

which establishes the Lipschitz constant for β. To conclude the proof, notice that

$$\begin{aligned} |f(h_2,\beta_2) - f(h_1,\beta_1)| &\leq |f(h_2,\beta_2) - f(h_2,\beta_1)| + |f(h_2,\beta_1) - f(h_1,\beta_1)| \\ &\leq 2\|A\| n |\beta_2 - \beta_1| + 2\|A\|\sqrt{n}\|h_2 - h_1\| \\ &= 2\|A\| n d_1((h_2,\beta_2),(h_1,\beta_1)) \end{aligned}$$

□

Thus, we need to cover $\mathcal{H}_u \times [-B,B]$ with respect to the distance d_1. Denote by $N(\epsilon) = \mathcal{N}(\epsilon, \mathcal{H}_u \times [-B,B], d_1)$ the covering number of this set. We can simplify this expression by noticing that if we choose an $\epsilon/2$ cover of $[-B,B]$ w.r.t. the absolute value distance and an $\epsilon/2$ cover of $\mathcal{H}_u$ w.r.t. d, their product gives an ϵ cover for $[-B,B] \times \mathcal{H}_u$ w.r.t. d_1. Also, it is clear that we can choose an $\epsilon/2$ cover of $[-B,B]$ of size $4B/\epsilon$. Thus, we have that

$$N(\epsilon) \leq C\frac{1}{\epsilon}\mathcal{N}\left(\frac{\epsilon}{2}, \mathcal{H}_u, d\right) \leq C\frac{1}{\epsilon}\mathcal{N}\left(\frac{\epsilon}{4}, \mathcal{H}, d\right). \tag{27}$$

Taking the union bound over this ϵ-net, we get that with probability at least $1 - C\log n\, N(\epsilon)\, \mathrm{e}^{-u^2}$, for any h, β in the net, $\varphi(h,\beta) - \varphi(h^*,\beta^*) \geq cu^2$. Now we would like to show that this implies the claim for all h that are far apart, not just the ones in the net. Since φ is n-Lipschitz with respect to d_1, if we choose $\epsilon = u^2/(2n)$, then this implies that with probability at least $1 - C\log n\, N(\epsilon)\, \mathrm{e}^{-u^2}$, for any $(h,\beta) \in \mathcal{H}_u \times [-B,B]$,

$$\varphi(h,\beta) > \varphi(h^*,\beta^*).$$

In order for this to hold with probability $1-\delta$, we need

$$CN\left(\frac{u^2}{2n}\right)\log n \mathrm{e}^{-u^2} < \delta \implies u > C(M)\sqrt{\log\log n + \log N\left(\frac{u^2}{2n}\right) + \log\frac{1}{\delta}}$$

Since u will end up being an upper bound for $\psi(\hat{h},\hat{\beta})$, we should select

$$\inf\left\{u : u > C(M)\sqrt{\log\log n + \log N\left(\frac{u^2}{2n}\right) + \log\frac{1}{\delta}}\right\} \leq C\inf\left\{\sqrt{n\epsilon} : n\epsilon > \log\log n + \log N(\epsilon) + \log\frac{1}{\delta}\right\}$$

By continuity arguments, it is easy to see that the infimum of the latter expression is achieved when

$$n\epsilon = \log\log n + \log N(\epsilon) + \log\frac{1}{\delta}$$

Substituting the bound of $N(\epsilon)$, it is enough to consider

$$n\epsilon = \log\log n + \log\frac{1}{\epsilon} + \log\mathcal{N}(\epsilon, \mathcal{H}) + \log\frac{1}{\delta}$$

Now notice that the right hand side of this inequality will be > 1 for δ smaller than a constant. This means that $\epsilon > 1/n$, which implies that $\log(1/\epsilon) < \log n$. Hence, the critical value of ϵ can only increase if we solve instead

$$n\epsilon = \log n + \log\mathcal{N}(\epsilon, \mathcal{H}) + \log\frac{1}{\delta}$$

This critical value of ϵ is the same as in the problem

$$\inf_{\epsilon\geq 0}\left\{n\epsilon + \log n + \log\mathcal{N}(\epsilon, \mathcal{H}) + \log\frac{1}{\delta}\right\}$$

Hence, we conclude that

$$\inf\left\{\sqrt{n\epsilon} : n\epsilon > \log\log n + \log N(\epsilon) + \log\frac{1}{\delta}\right\} \leq \sqrt{\inf_{\epsilon\geq 0}\left\{n\epsilon + \log n + \log\mathcal{N}(\epsilon, \mathcal{H}) + \log\frac{1}{\delta}\right\}}$$

Let $\epsilon*$ be the point where the infimum is reached. Then, with probability at least $1-\delta$, for all $(h, \beta) \in \mathcal{H}_{u^*} \times [-B, B]$, with $u = \sqrt{n\epsilon^*}$, we have

$$\varphi(h, \beta) > \varphi(h^*, \beta^*)$$

However, we know by definition of MPLE that $\varphi(\hat{h}, \hat{\beta}) \leq \varphi(h^*, \beta^*)$, which implies that with probability $1-\delta$,

$$\psi(\hat{h}, \hat{\beta}) \leq C(M)(u^*)^2 = C(M)\sqrt{n\epsilon^*} = C(M)\sqrt{\inf_{\epsilon\geq 0}\left\{n\epsilon + \log\mathcal{N}(\epsilon, \mathcal{H}) + \log\frac{n}{\delta}\right\}}$$

Now we use inequality (27) and the proof is complete.

□

We are now ready to complete the proof of Theorem 7. So far, we have proved that for the MPLE estimates we have that $\psi(\hat{h}, \hat{\beta})$ will be small. However, we still need to show that this implies that the estimation of the parameters will be good enough. Thus, we now present the complete proof, which shows exactly the connection of ψ with the estimation error of the parameters.

Proof of Theorem 7. Assume that the high probability event of Lemma 10 holds and let R denote the right hand side of (26). Our goal is to prove that $\|\hat{h} - h^*\|_2^2/n \leq R\mathcal{C}_1(\mathcal{H}, h^*, \beta^*)$. We can assume that $R \geq 1$, and we divide into cases. First, if $\psi(\hat{h}, \hat{\beta}) \leq 1$, then by definition of $\mathcal{C}_1$ in (15),

$$\begin{aligned}\frac{\|\hat{h} - h^*\|_2^2}{n} &= \min\left(\frac{\|\hat{h} - h^*\|_2^2/n}{\psi(\hat{h}, \hat{\beta})}, \frac{\|\hat{h} - h^*\|_2^2}{n}\right) \\ &\leq \mathcal{C}_1(\mathcal{H}, h^*, \beta^*) \leq R\mathcal{C}_1(\mathcal{H}, h^*, \beta^*).\end{aligned}$$

Otherwise,

$$\frac{\|\hat{h} - h^*\|_2^2}{n} = \psi(\hat{h}, \hat{\beta}) \cdot \frac{\|\hat{h} - h^*\|_2^2/n}{\psi(\hat{h}, \hat{\beta})} \leq R \cdot \mathcal{C}_1(\mathcal{H}, h^*, \beta^*). \tag{28}$$

Similarly, we can bound $(\hat{\beta} - \beta^*)^2$ in terms of $\mathcal{C}_2$.

□

B.2 Exchangeable pairs

We use the following theorem, proven in [Cha05], that guarantees concentration of a function $f(X)$, of a random variable X:

Theorem 9. *Let X be a random variable over some domain $\mathcal{X}$ and let $f\colon \mathcal{X} \to \mathbb{R}$ be a measurable map. Let X' be another $\mathcal{X}$-valued random variable, jointly distributed with X, such that (X, X') is an exchangeable pair, namely, (X, X') has the same distribution as (X', X). Let $F\colon \mathcal{X}^2 \to \mathbb{R}$ be a measurable map that is antisymmetric, namely,*

$$F(x, x') - F(x', x),$$

and further

$$\mathbb{E}[F(X, X')|X] = F(X).$$

Define for $x \in X$,

$$v(x) = \frac{1}{2}\mathbb{E}[|(f(X) - f(X'))F(X, X')| \mid X = x].$$

Let $M > 0$ and assume that $|v(X)| \leq M$ almost surely. Then, for all $t > 0$,

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| > t] \leq 2\mathrm{e}^{-t^2/(2M)}.$$

We prove Lemma 8:

Proof of Lemma 8. First, notice that $\mathbb{E}f(\sigma) = 0$. Indeed, let σ_{-i} denote the vector obtained from σ by omitting σ_i and notice that $\mathbb{E}[\sigma_i \mid \sigma_{-i}] = \tanh(\beta^*(A\sigma)_i + h_i^*)$. Taking expectation, one obtains that

$$\mathbb{E}[\sigma_i - \tanh(\beta^*(A\sigma)_i + h_i^*)] = \mathbb{E}[\mathbb{E}[\sigma_i \mid \sigma_{-i}] - \tanh(\beta^*(A\sigma)_i + h_i^*)] = 0,$$

which implies that $\mathbb{E}f(\sigma) = 0$ as required.

Next, we prove concentration around the expectation. Define σ' in a joint distribution with σ as follows: given σ, an index $j \in [n]$ is drawn uniformly at random. Then, σ' is obtained by $\sigma'_i = \sigma_i$ for $i \neq j$ and σ'_i is drawn from the conditional distribution of σ_i conditioned on σ_{-i}. One can indeed verify that (σ, σ') is an exchangeable pair.

Define the function:

$$F(\sigma, \sigma') = n\sum_{i=1}^{n} v_i(\sigma_i - \sigma'_i)$$

and notice that F is antisymmetric. Further notice that

$$\mathbb{E}[F(\sigma, \sigma') \mid \sigma, j] = nv_j(\sigma_j - \mathbb{E}[\sigma'_j \mid \sigma, j]) = nv_j(\sigma_j - \mathbb{E}[\sigma_j \mid \sigma_{-j}]) = nv_j(\sigma_j - \tanh(\beta^*(A\sigma)_i + h_i^*)).$$

Taking expectation over j, one derives that $\mathbb{E}[F(\sigma, \sigma') \mid \sigma] = f(\sigma)$ as required.

Lastly, we bound $v(\sigma) = \mathbb{E}[|(f(\sigma) - f(\sigma'))F(\sigma, \sigma')| \mid \sigma]$. One has

$$v(\sigma) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[|(f(\sigma) - f(\sigma'))F(\sigma, \sigma')| \mid \sigma, j].$$

Condition on j. Then,

$$|F(\sigma, \sigma')| = n|v_j(\sigma_j - \sigma'_j)| \leq 2nv_j.$$

Further, using the 1-Lipschitzness of tanh,

$$|f(\sigma) - f(\sigma')| \leq |v_j(\sigma_j - \sigma'_j)| + \sum_{i=1}^{n} |v_i(\tanh(\beta^*(A\sigma)_i + h_i^*) - \tanh(\beta^*(A\sigma')_i + h_i^*))|$$

$$\leq 2|v_j| + \sum_{i=1}^{n} |v_i \beta^* ((A\sigma)_i - (A\sigma')_i)| = 2|v_j| + \sum_{i=1}^{n} |v_i \beta^* A_{ij} (\sigma_j - \sigma'_j)|$$
$$\leq 2|v_j| + 2\beta^* \sum_{i=1}^{n} |v_i||A_{ij}|.$$

We derive that, conditioned on j, one has

$$|F(\sigma, \sigma')(f(\sigma) - f(\sigma'))| \leq 4n|v_j| \left(|v_j| + \beta^* \sum_{i=1}^{n} |v_i||A_{ij}| \right).$$

Hence, for all σ,

$$v(\sigma) = \mathbb{E}[|F(\sigma, \sigma')(f(\sigma) - f(\sigma'))| \mid \sigma] = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[|F(\sigma, \sigma')(f(\sigma) - f(\sigma'))| \mid \sigma, j]$$
$$\leq \frac{1}{n} \sum_{j=1}^{n} 4n|v_j| \left(|v_j| + |\beta^*| \sum_{i=1}^{n} |v_i||A_{ij}| \right).$$

Denote $\tilde{v} = (|v_1|, \ldots, |v_n|)$ and similarly, $\tilde{A} = (|A_i j|)_{i,j \in [n]}$, and we derive that

$$v(\sigma) \leq 4\tilde{v}^\top \tilde{v} + 4|\beta^*| \tilde{v}^\top \tilde{A} \tilde{v} \leq 4\|\tilde{v}\|_2^2 + 4|\beta^*| \|\tilde{v}\|_2 \|\tilde{A}\|_2 \|\tilde{v}\|_2 \leq 4\|\tilde{v}\|_2^2 \left(1 + |\beta^*| \|\tilde{A}\|_\infty\right) = 4\|v\|_2^2 \left(1 + |\beta^*| \|A\|_\infty\right).$$

Applying Theorem 9 with $M = 4\|v\|_2^2 (1 + |\beta^*| \|A\|_\infty)$, the result follows. □

B.3 Proofs of Auxiliary lemmas

The proofs below follow from similar arguments as in [Dag+20]. Below, we elaborate on how to modify the arguments in that paper.

B.3.1 Proof of Lemma 3

We use the index sets $I_1, \ldots, I_\ell \subseteq [n]$ created in [Dag+20], where $\ell = O(\log n)$. We start by bounding the first derivative of φ. The following is a small modification of Lemma 3 in [Dag+20]:

Lemma 13. *For any $h \in \mathbb{R}^n$, $\beta \in \mathbb{R}$ and $t > 0$ we have*

$$\left| (h, \beta)^\top \nabla \varphi(\beta^*, h^*) \right| \leq C \max \left(t \left(\|\beta A\|_F + \max_{j \in [l]} \|\mathbb{E}[\beta A \sigma + h | \sigma_{-I_j}]\|_2 \right), t^2 \|\beta A\|_2 \right)$$

with probability at least

$$1 - C \log n \, \exp\left(-ct^2\right).$$

Proof. This follows from a similar proof as of the proof of Lemma 3 in [Dag+20]. We note the differences. First, we replace $\varphi(J)$ with $\varphi(h, \beta)$. Second, we replace $\psi_j(A; \sigma) = \left| \sum_{i \in I_j} A_i \sigma (\sigma_i - \tanh(J_i^* \sigma)) \right|$ with $\psi_i(h, \beta; \sigma) = \left| \sum_{i \in I_j} (\beta A_i \sigma + h_i)(\sigma_i - \tanh(\beta^* A_i x)) \right|$. Thirdly, in the proof of Lemma 3 in [Dag+20], we replace the inequality

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

with

$$\mathbb{P}\left[|\psi_j(h, \beta; \sigma)| \geq \max\left(t \left(\|\beta A\|_F + \max_{j \in [\ell]} \|\mathbb{E}[\beta A \sigma + h | \sigma_{-I_j}]\|_2 \right), t^2 \|\beta A\|_2 \right) \right] \leq \exp\left(-ct^2\right),$$

which follows from Lemma 5 in [Dag+20]. □

Next, we bound the second derivative of φ. We use Lemma 4 which claims that for all $h' \in [-M, M]^n$ and $\beta' \in [-B, B]$,

$$(h, \beta)^\top \nabla^2 \varphi(h', \beta')(h, \beta) \geq c\|\beta A\sigma + h\|_2$$

and bound $\|\beta A\sigma + h\|_2$. We state a modified variant of Lemma 4 in [Dag+20]:

Lemma 14. *For any $\beta \in \mathbb{R}$, $h \in \mathbb{R}^n$ and $t > 0$, with probability at least $1 - \log n e^{-ct^2}$ we have that either*

$$\|\beta A\sigma + h\|_2^2 \geq c'\|\beta A\|_F^2 + c' \max_{j \in [\ell]} \|\mathbb{E}[\beta A\sigma + h \mid \sigma_{-I_j}]\|_2^2, \tag{29}$$

or, that

$$\|\beta A\|_F + \max_{j \in [\ell]} \|\mathbb{E}[\beta A\sigma + h \mid \sigma_{-I_j}]\|_2 \leq Ct\|\beta A\|_2. \tag{30}$$

Proof. We explain the changes with respect to the proof of Lemma 4 in [Dag+20].

- We start by replacing Lemma 7 in [Dag+20] with

$$\mathbb{E}\left[\|\beta A\sigma + h\|_2^2 \;\middle|\; \sigma_{-I_j}\right] \geq c_e\|\beta A_{\cdot I_j}\|_F^2 + \left\|\mathbb{E}\left[\beta A\sigma + h \;\middle|\; \sigma_{-I_j}\right]\right\|_2^2, \tag{31}$$

  where $c_e > 0$ is a constant.

- Next, we replace Lemma 8 in [Dag+20] with the claim that for any $t > 0$,

$$\mathbb{P}\left[\|\beta A\sigma + h\|_2^2 < \mathbb{E}\left[\|\beta A\sigma + h\|_2^2 \;\middle|\; \sigma_{-I_j}\right] - t \;\middle|\; \sigma_{-I_j}\right] \leq \exp\left(-\frac{c}{\|\beta A\|_2^2} \min\left(\frac{t^2}{\|\beta A\|_F^2 + \|\mathbb{E}\left[\beta A\sigma + h \;\middle|\; \sigma_{-I_j}\right]\|_2^2}, t\right)\right). \tag{32}$$

  This follows from the same proof as that of Lemma 8 in [Dag+20].

- Finally, we explain how to derive Lemma 14. For any $j \in [\ell]$, let E_j denote the event that

$$\|\beta A\sigma + h\|_2^2 > c_e\|\beta A_{\cdot I_j}\|_F^2 + \left\|\mathbb{E}\left[\beta A\sigma + h \;\middle|\; \sigma_{-I_j}\right]\right\|_2^2 - t\|\beta A\|_2 \left(\|\beta A\|_F + \|\mathbb{E}[\beta A\sigma + h \mid \sigma_{-I_j}]\|_2\right),$$

  where c_e is the constant from (31). By (32),

$$\mathbb{P}[E_j \mid \sigma_{-I_j}] \geq 1 - \exp\left(-c\min\left(t^2, t(\|\beta A\|_F + \|\mathbb{E}[\beta A\sigma + h \mid \sigma_{-I_j}]\|_2)/\|\beta A\|_2\right)\right) \geq 1 - \exp\left(-ct^2\right),$$

  using the fact that E_j holds trivially for $t \geq \Omega((\|\beta A\|_F + \|\mathbb{E}[\beta A\sigma + h \mid \sigma_{-I_j}]\|_2)/\|\beta A\|_2)$. We conclude that

$$\mathbb{P}[E_j] \geq 1 - \exp(-ct^2).$$

  Taking a union bound over $j \in [\ell]$ and that $\ell = O(\log n)$, one has

$$\mathbb{P}[\cap_j E_j] \geq 1 - C\log(n)\exp(-ct^2).$$

  Assume that $\cap_j E_j$ holds. Then, one can easily verify that if (30) does not hold then (29) holds, if the constants in (29) and (30) are chosen appropriately.

□

Proof of Lemma 3. Let $t \geq 0$ and assume that the high probability events of both Lemma 13 and Lemma 14 hold. As discussed above, due to Lemma 4 it suffices to show that under these high probability events, either

$$\left|(h, \beta)^\top \varphi(h^*, \beta^*)\right| / \|\beta A\sigma + h\|_2^2 \leq \frac{C't}{\|\beta A\sigma + h\|_2} \tag{33}$$

holds or

$$\|\beta A\sigma + h\|_2 \leq Ct\|\beta A\|_2.$$

We divide into cases.

- First, assume that
$$\|\beta A\|_F + \max_{j\in[l]} \|\mathbb{E}[\beta A\sigma + h \mid \sigma_{-I_j}]\|_2 \le Ct\|\beta A\|_2$$
for a sufficiently large constant $C > 0$. Then, by Lemma 13,
$$\left|(h,\beta)^\top \varphi(h^*,\beta^*)\right| \le t^2\|\beta A\|_2.$$
This implies that if $\|\beta A\sigma + h\|_2 \ge t\|\beta A\|_2$ then (33) holds, which is what we wanted to prove.

- Next, assume that
$$\|\beta A\|_F + \max_{j\in[l]} \|\mathbb{E}[\beta A\sigma + h \mid \sigma_{-I_j}]\|_2 \ge Ct\|\beta A\|_2.$$
Since we assumed that the high probability events of Lemma 13 and Lemma 14 hold, one has that
$$\left|(h,\beta)^\top \varphi(h^*,\beta^*)\right| \le Ct\left(\|\beta A\|_F + \max_{j\in[l]} \|\mathbb{E}[\beta A\sigma + h|\sigma_{-I_j}]\|_2\right),$$
whereas
$$\|\beta A\sigma + h\|_2^2 \ge c'\|\beta A\|_F^2 + c' \max_{j\in[l]} \|\mathbb{E}[\beta A\sigma + h \mid \sigma_{-I_j}]\|_2^2,$$
which implies that
$$\left|(h,\beta)^\top \varphi(h^*,\beta^*)\right| \le C't\|\beta A\sigma + h\|,$$
which derives (33) as required.

□

B.3.2 Proof of Lemma 4

This holds from a simple calculation as Equation (15) in [Dag+20].

B.3.3 Proof of Lemma 6

This follows directly from the following lemma:

Lemma 15. *For any real matrix A and any $h \in \mathbb{R}^n$,*
$$\mathbb{P}[\|A\sigma + h\|_2 \ge c\|A\|_F] \ge 1 - \exp(-c\|A\|_F^2/\|A\|_2^2).$$

Sketch. This is analogous to the second part of Lemma 4 in [Dag+20]. □

B.4 Proof of Lemma 1

Notice the following: first,
$$\sqrt{\psi(h,\beta;h^*,\beta^*)} \ge |\beta - \beta'|\|A\|_F$$
and secondly,
$$\begin{aligned}\sqrt{\psi(h,\beta;h^*,\beta^*)} &\ge \left\|h - h^* + (\beta-\beta^*)A\mathbf{tanh}\left(\frac{\beta^*}{\beta-\beta^*}(h^*-h) + h^*\right)\right\| \\ &\ge \|h-h^*\| - |\beta-\beta^*|\|A\|_2\left\|\mathbf{tanh}\left(\frac{\beta'}{\beta-\beta'}(h^*-h) + h^*\right)\right\|_2 \\ &\ge \|h-h^*\| - |\beta-\beta^*|\|A\|_2\sqrt{n} \ge \|h-h^*\| - |\beta-\beta^*|\sqrt{n}\ ,\end{aligned}$$
using the fact that $|\tanh(a)| \le 1$ for all $a \in \mathbb{R}$ and that $\|A\|_2 \le \|A\|_\infty \le 1$, where the first inequality is due to the symmetricity of A and the second is due to the assumption in this theorem that $\|A\|_\infty \le 1$.

Divide into cases: if

$$\sqrt{n}|\beta - \beta^*| \le \|h - h^*\|_2,$$

then

$$\sqrt{\psi(h, \beta; h^*, \beta^*)} \ge \|h - h^*\|/2,$$

which implies that

$$\frac{\|h - h^*\|^2/n}{\psi(h, \beta; h^*, \beta^*)} \ge \frac{1}{4n} \ge \frac{1}{4\|A\|_F^2},$$

using the fact that $\|A\|_F \le \sqrt{n}\|A\|_2 \le \sqrt{n}$. For the second case, if

$$\sqrt{n}|\beta - \beta^*| \ge \|h - h^*\|/2,$$

then

$$\sqrt{\psi(h, \beta; h^*, \beta^*)} \ge |\beta - \beta^*|\|A\|_F \ge \|h - h^*\|_2\|A\|_F/2\sqrt{n}.$$

Hence,

$$\frac{\|h - h^*\|^2/n}{\psi(h, \beta; h^*, \beta^*)} \ge \frac{1}{4\|A\|_F^2}.$$

This holds for any h, β, hence

$$\mathcal{C}_1(\mathcal{H}, h^*, \beta^*) \le \mathcal{C}_1'(\mathcal{H}, h^*, \beta^*) \le \frac{1}{4\|A\|_F^2},$$

as required.

C Lower bound

We start with proving the lower bound in Theorem 6 and then prove Theorem 3.

C.1 Proof of the lower bound in Theorem 6

As stated in Section A, it suffices to prove Theorem 8. The first step is to find a suitable upper bound for the KL-divergence between two Ising models in high temperature. Ideally, we would like this bound to depend on the difference in β and the difference in h. This is achieved with the following Lemma.

Lemma 16. *Suppose $h_0, h_1, \beta_0, \beta_1$ are such that $\|h_0\|_\infty, \|h_1\|_\infty \le M$ and $|\beta_0|\|A\|_\infty, |\beta_1|\|A\|_\infty \le 1 - \alpha$ for some $\alpha \in (0, 1)$. Then, there is a constant $C = C(\alpha, M)$ such that*

$$D_{KL}(P_{h_0,\beta_0}\|P_{h_1,\beta_1}) \le C(\beta_1 - \beta_0)^2\|A\|_F^2 + \|(\beta_1 - \beta_0)\mathbb{E}_{h_0,\beta_0}[A\sigma] + h_1 - h_0\|_2^2.$$

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

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

which is the log partition function of a model with interaction matrix βA and external field h. Also, we also define for $t \in [0, 1]$,

$$h_t = (1 - t)h_0 + th_1; \quad \beta_t = (1 - t)\beta_0 + t\beta_1$$

and notice that for $t \in \{0, 1\}$, this definition coincides with $h_0, h_1, \beta_0, \beta_1$. Some simple calculations show that

$$\frac{dF(h_t, \beta_t)}{dt} = \mathbb{E}_{h_t,\beta_t}\left[\frac{1}{2}(\beta_1 - \beta_0)\sigma^\top A\sigma + \sigma^\top(h_1 - h_0)\right]$$

$$\frac{d^2F(h_t,\beta_t)}{dt^2} = \mathrm{Var}_{h_t,\beta_t}\left[\frac{1}{2}(\beta_1-\beta_0)\sigma^\top A\sigma + \sigma^\top(h_1-h_0)\right]$$

where the expectation and variance are over $\sigma \sim \mathbb{P}_{h_t,\beta_t}$ Now, we do some simple calculations with the KL-divergence

$$\begin{aligned}
D_{KL}(P_{h_0,\beta_0}\|P_{h_1,\beta_1}) &= \mathbb{E}_{h_0,\beta_0}\ln\frac{P_{h_0,\beta_0}(\sigma)}{P_{h_1,\beta_1}(\sigma)} \\
&= \mathbb{E}_{h_0,\beta_0}\left[\frac{1}{2}(\beta_0-\beta_1)\sigma^\top Ay + \sigma^\top(h_0-h_1) - F(h_0,\beta_0) + F(h_1,\beta_1)\right] \\
&= F(h_1,\beta_1) - F(h_0,\beta_0) - \frac{dF(h_t,\beta_t)}{d_t}\Big|_{t=0} \\
&= \frac{1}{2}\frac{d^2F(h_t,\beta_t)}{d_t^2}\Big|_{t=\xi} \\
&= \mathrm{Var}_{h_\xi,\beta_\xi}\left[\frac{1}{2}(\beta_1-\beta_0)\sigma^\top A\sigma + \sigma^\top(h_1-h_0)\right]
\end{aligned} \tag{34}$$

for some $\xi \in [0,1]$, using Taylor's approximation on $F(h_t,\beta_t)$. Hence, it suffice to bound this variance in order to obtain a bound on the KL-divergence. This is a variance of a quadratic polynomial of the Ising model, which can be bounded by the following theorem:

Theorem 10 ([Ada+19], Theorem 2.2)**.** *Let A be an interaction matrix of dimension $n\times n$, let $\beta\in\mathbb{R}$ such that $\|\beta A\|_\infty \le 1-\alpha$ for some $\alpha\in(0,1)$ and let $h\in[-M,M]^n$ for some $M>0$. Let A' be an $n\times n$ real matrix and let $h'\in\mathbb{R}^n$. Then, there exists $C=C(\alpha,M)>0$ such that*

$$\mathrm{Var}_{h,\beta}\left[\frac{1}{2}\sigma^\top A'\sigma + \sigma^\top h\right] \le C\|A'\|_F^2 + C\|\mathbb{E}_{h,\beta}[A\sigma+h]\|_2^2.$$

We can apply this theorem to bound the right hand side of (34), since the assumptions of this lemma guarantee that $\|\beta_t A\|_\infty \le \max(\|\beta_0 A\|_\infty, \|\beta_1 A\|_\infty) \le 1-\alpha$, and the result follows. □

Now we will try to connect this bound on the KL-divergence with the rate of the upper bound. In the proof of Theorem 7, in order to obtain a good upper bound on the rate, we lower bounded $\|h^*-h-(\beta-\beta^*)A\sigma\|_2^2$ by a nonrandom quantity(Lemma 5. In the following Lemma, we essentially show that this substitution is tight in the high temperature regime, which means there is a corresponding upper bound of the same order. The proof involves a similar trick as in Lemma 5, by bounding each term of the triangle inequality separately.

Lemma 17. *Suppose we have an Ising model with β^*, h^* satisfying $\beta^*\|A\|_\infty \le 1-\alpha$ for some $\alpha\ge 0$ and let h,β such that $|\beta|,\|h\|_\infty \le M$. Then, there exists some $C=C(\alpha,M)>0$, such that*

$$\|h^*-h-(\beta-\beta^*)\mathbb{E}_{h^*,\beta^*}A\sigma\| \le C\left\|h^*-h-(\beta-\beta^*)A\tanh\left(\frac{\beta^*}{\beta-\beta^*}(h^*-h)+h^*\right)\right\| + C\|(\beta-\beta^*)A\|_F.$$

Proof. Denote $\tilde\beta = \beta-\beta^*$. We have

$$\begin{aligned}
&\|h^*-h-(\beta-\beta^*)\mathbb{E}A\sigma\| \le \\
&\left\|h^*-h-\tilde\beta A\tanh\left(\frac{\beta^*}{\tilde\beta}(h^*-h)+h^*\right)\right\| + \left\|\tilde\beta A\tanh(\beta^*\mathbb{E}[A\sigma]+h^*) - \tilde\beta A\tanh\left(\frac{\beta^*}{\tilde\beta}(h^*-h)+h^*\right)\right\| \\
&\qquad + \left\|\tilde\beta\mathbb{E}\left[A\tanh(\beta^*A\sigma+h^*)\right] - \tilde\beta A\tanh(\beta^*\mathbb{E}[A\sigma]+h^*)\right\|
\end{aligned} \tag{35}$$

We bound the right hand side of (35). The second term is bounded as follows, using the fact that tanh is a 1-Lipschitz function:

$$\begin{aligned}
&\left\|\tilde{\beta}A\tanh(\beta^*\mathbb{E}[A\sigma]+h^*)-\tilde{\beta}A\tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^*-h)+h^*\right)\right\|_2 \\
&\qquad\le|\tilde{\beta}|\|A\|_2\left\|\tanh(\beta^*\mathbb{E}[A\sigma]+h^*)-\tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^*-h)+h^*\right)\right\|_2 \\
&\qquad\le|\tilde{\beta}|\|A\|_2\left\|\beta^*\mathbb{E}[A\sigma]+h^*-\left(\frac{\beta^*}{\tilde{\beta}}(h^*-h)+h^*\right)\right\|_2 \\
&\qquad=|\beta^*|\|A\|_2\left\|\tilde{\beta}\mathbb{E}[A\sigma]-(h^*-h)\right\|_2=|\beta^*|\|A\|_2\left\|h^*-h-\tilde{\beta}\mathbb{E}[A\sigma]\right\|_2.
\end{aligned}$$

Further, the last term of (35) can be bounded as follows, using Jenssen's inequality and the Lipschitzness of tanh:

$$\begin{aligned}
&\left\|\tilde{\beta}\mathbb{E}\left[A\tanh(\beta^*A\sigma+h^*)\right]-\tilde{\beta}A\tanh(\beta^*\mathbb{E}[A\sigma]+h^*)\right\| \\
&\qquad\le|\tilde{\beta}|\|A\|_2\left\|\mathbb{E}\left[\tanh(\beta^*A\sigma+h^*)\right]-\tanh(\beta^*\mathbb{E}[A\sigma]+h^*)\right\|_2 \\
&\qquad=|\tilde{\beta}|\|A\|_2\sqrt{\left\|\mathbb{E}\left[\tanh(\beta^*A\sigma+h^*)-\tanh(\beta^*\mathbb{E}[A\sigma]+h^*)\right]\right\|_2^2} \\
&\qquad\le|\tilde{\beta}|\|A\|_2\sqrt{\mathbb{E}\left[\left\|\tanh(\beta^*A\sigma+h^*)-\tanh(\beta^*\mathbb{E}[A\sigma]+h^*)\right\|_2^2\right]} \\
&\qquad\le|\tilde{\beta}|\|A\|_2\sqrt{\mathbb{E}\left[\left\|\beta^*A\sigma+h^*-(\beta^*\mathbb{E}[A\sigma]+h^*)\right\|_2^2\right]} \\
&\qquad=|\tilde{\beta}|\|A\|_2|\beta^*|\sqrt{\mathbb{E}\left[\left\|A\sigma-\mathbb{E}[A\sigma]\right\|_2^2\right]}=|\tilde{\beta}|\|A\|_2|\beta^*|\sqrt{\sum_{i=1}^n\mathrm{Var}(A_i\sigma)}, \quad (36)
\end{aligned}$$

where A_i is row i of A. Using Theorem 10, each variance is bounded by $C\|A_i\|_2^2$, and the right hand side of (36) is bounded by

$$C|\tilde{\beta}|\|A\|_2|\beta^*|\sqrt{\sum_i\|A_i\|_2^2}\le C|\tilde{\beta}|\|A\|_2|\beta^*|\|A\|_F\le C|\tilde{\beta}|\|A\|_F.$$

since $\|A_2\||\beta^*|\le 1$. Combining the above calculations, we derive that

$$\|h^*-h-\tilde{\beta}\mathbb{E}A\sigma\|\le\left\|h^*-h-\tilde{\beta}A\tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^*-h)+h^*\right)\right\|+|\beta^*|\|A\|_2\|h^*-h-\tilde{\beta}\mathbb{E}A\sigma\|+C|\tilde{\beta}|\|A\|_F.$$

Since $|\beta^*|\|A\|_2<1$, we derive that

$$\|h^*-h-\tilde{\beta}\mathbb{E}A\sigma\|\le\frac{C}{1-|\beta^*|\|A\|_2}\left(\left\|h^*-h-\tilde{\beta}A\tanh\left(\frac{\beta^*}{\tilde{\beta}}(h^*-h)+h^*\right)\right\|+|\tilde{\beta}|\|A\|_F\right)$$

□

As a corollary of the two lemmas above, we derive a bound for the KL-divergence between two Ising models in terms of a familiar quantity that was used in the upper bound of the rate.

Lemma 18. *Suppose we have an Ising model with β_0, h_0 satisfying $\beta_0\|A\|_\infty\le 1-\alpha$ for some $\alpha\ge 0$ and let h_1,β_1 such that $|\beta_1|,\|h_1\|_\infty\le M$. Then, there exists some $C=C(\alpha,M)>0$ such that*

$$\begin{aligned}
D_{KL}(P_{h_0,\beta_0}\|P_{h_1,\beta_1})&\le C\left\|h_0-h_1-(\beta_1-\beta_0)A\tanh\left(\frac{\beta_0}{\beta_1-\beta_0}(h_0-h_1)+h_0\right)\right\|_2^2+C\|(\beta_1-\beta_0)A\|_F^2 \\
&=C\psi(h_1,\beta_1;h_0,\beta_0).
\end{aligned}$$

Next, we add another auxiliary lemma:

Lemma 19. *Let $\beta_0, \beta_1 \in \mathbb{R}$ and $h_0, h_1 \in \mathbb{R}^n$. Define for any $t \in \mathbb{R}$,*

$$(h_t, \beta_t) = (h_0, \beta_0) + t(h_1 - h_0, \beta_1 - \beta_0).$$

Then, for any $t \in \mathbb{R}$,

$$\psi(h_t, \beta_t; h_0, \beta_0) = t^2 \psi(h_1, \beta_1; h_0, \beta_0).$$

Consequently, for any $t \neq 0$,

$$\frac{\|h_t - h_0\|^2}{\psi(h_t, \beta_t; h_0, \beta_0)} = \frac{\|h_1 - h_0\|^2}{\psi(h_1, \beta_1; h_0, \beta_0)}.$$

Proof. This follows from the fact that

$$h_t - h_0 = t(h_1 - h_0); \quad \beta_t - \beta_0 = t(\beta_1 - \beta_0).$$

□

To prove the lower bound, we use the following version of Le-Cam's method for binary hypothesis testing:

Lemma 20 (Le-Cam)**.** *Let (Θ, d) be a metric space, let $\{P_\theta : \theta \in \Theta\}$ be a family of distributions over the domain $\mathcal{X}$ let $\theta_0, \theta_1 \in \Theta$ and let $\hat{\theta} : \mathcal{X} \to \Theta$ be some estimator. Then, there exists $i \in \{0, 1\}$ such that*

$$\mathbb{P}_{x \sim P_{\theta_i}}[d(\hat{\theta}(x), \theta_i) \geq d(\theta_0, \theta_1)/2] \geq \frac{1 - D_{TV}(P_{\theta_0}, P_{\theta_1})}{2}.$$

We are now ready to present the proof of the Lower bound in Theorem 6, which is also the proof of Theorem 8. The trick is to use the bound of Lemma 18 to select to Ising models that are as far as possible, while having KL-divergence bounded by a constant. Using Pinsker's inequality, this translates into a constant bound for the TV-distance between them, which implies that we cannot distinguish between the two models.

Proof. Let (h_0, β_0) be the maximizer of $\mathcal{C}_1(\mathcal{H}, h, \beta)$ over $(h, \beta) \in \mathcal{H} \times [-1/2, 1/2]$. Let (h_1, β_1) be the maximizer in the definition of $\mathcal{C}_1(\mathcal{H}, h_0, \beta_0)$. Let $c_0 \in (0, 1)$ be a constant to be defined later. Define (h_t, β_t) as in Lemma 19. And let ξ be the maximal number, $t \in [0, 1]$ such that the following holds: both

$$t^2 \psi(h_1, \beta_1; h_0, \beta_0) = \psi(h_t, \beta_t; h_0, \beta_0) \leq c_0, \tag{37}$$

and

$$|\beta_t - \beta_0| \leq 1/2. \tag{38}$$

Notice that there exists $\zeta \in \{-\xi, \xi\}$ such that $|\beta_\zeta| \leq 1/2$: indeed, $|\beta_0| \leq 1/2$, and $-1/2 \leq \beta_0 - \beta_\xi = -(\beta_0 - \beta_{-\xi}) \leq 1/2$. Fix such ζ and notice that due to the assumption that $\mathcal{F}$ is convex and closed under complements, we have that $h_\zeta \in \mathcal{F}$. Furthermore, by Lemma 18,

$$D_{KL}(\mathbb{P}_{h_\zeta, \beta_\zeta} \parallel \mathbb{P}_{h_0, \beta_0}) \leq C\psi(h_\zeta, \beta_\zeta; h_0, \beta_0) \leq Cc_0$$

which implies by Pinsker's inequality, that

$$D_{TV}(\mathbb{P}_{h_\zeta, \beta_\zeta}, \mathbb{P}_{h_0, \beta_0}) \leq \sqrt{D_{KL}(\mathbb{P}_{h_\zeta, \beta_\zeta} \parallel \mathbb{P}_{h_0, \beta_0})/2} \leq \sqrt{Cc_0/2}.$$

By Le-Cam's method, if c_0 is sufficiently small, then either

$$\mathbb{P}_{h_\zeta, \beta_\zeta}[\|\hat{h} - h_\zeta\| \geq \|h_\zeta - h_0\|/2] \geq 0.49$$

or

$$\mathbb{P}_{h_0, \beta_0}[\|\hat{h} - h_0\| \geq \|h_\zeta - h_0\|/2] \geq 0.49.$$

It remains to prove that

$$\|h_\zeta - h_0\|^2/n \geq c\mathcal{C}_1(\mathcal{H}, h_0, \beta_0) \geq c\mathcal{C}_1(\mathcal{H}, h_\zeta, \beta_\zeta)$$

for some constant $c > 0$. Notice that the second inequality follows from the fact that h_0, β_0 was chosen to maximize $\mathcal{C}_1$, hence it remains to prove the first inequality. By Lemma 19, one has

$$\frac{\|h_\zeta - h_0\|^2}{\psi(h_\zeta, \beta_\zeta; h_0, \beta_0)} = \frac{\|h_1 - h_0\|^2}{\psi(h_1, \beta_1; h_0, \beta_0)}.$$

If $\psi(h_1, \beta_1; h_0, \beta_0) \leq 1$, then $\mathcal{C}_1(\mathcal{H}, h_0, \beta_0) = \|h_0 - h_1\|_2^2/n$, by definition of $\mathcal{C}_1$ and by the fact (h_1, β_1) was chosen as the maximizer in its definition. Since $\psi(h_1, \beta_1; h_0, \beta_0) \leq 1$, we can choose ζ to be at least a constant c, without violating the conditions(37), (38). Hence

$$\|h_\zeta - h_0\|^2/n \geq c^2\|h_1 - h_0\|^2/n = c^2\mathcal{C}_1(\mathcal{H}, h_0, \beta_0)$$

as required. Next, assume that $\psi(h_1, \beta_1; h_0, \beta_0) > 1$, which implies that

$$\mathcal{C}_1(\mathcal{H}, h_0, \beta_0) = \frac{\|h_0 - h_1\|_2^2/n}{\psi(h_1, \beta_1; h_0, \beta_0)} = \frac{\|h_0 - h_\zeta\|_2^2/n}{\psi(h_\zeta, \beta_\zeta; h_0, \beta_0)}.$$

By definition of ζ, we derive that $\psi(h_\zeta, \beta_\zeta; h_0, \beta_0) \geq c$ for some constant $c \geq 0$. This happens because we can take ζ at least a constant without violating (38), which will result in a greater than constant value for $\psi(h_\zeta, \beta_\zeta; h_0, \beta_0)$. In particular, $\|h_0 - h_\zeta\|_2^2/n \geq c\mathcal{C}_1(\mathcal{H}, h_0, \beta_0)$ as required. This concludes the proof. □

C.2 Proof of Theorem 3

It suffices to prove the statement assuming that n and r are integer powers of 2, and we make this assumption here. Further, we can replace the assumption that $|x_i| \leq M$ and $|\theta| \leq 1$ with $|x_i| \leq 1$ and $|\theta| \leq M$. Here, we will prove the result with $M = 2$.

Define $a = 0.8952...$ as the non-negative solution to $\tanh(1 + a/2) = a$. Consider the logistic regression setting where $x = \mathbf{1} = (1, \ldots, 1)$. Let $\theta_0 = 1$, $\beta_0 = 1/2$, $\theta_1 = 1 + a$ and $\beta_1 = -1/2$. Let A be the following matrix: partition the indices to r sets of size n/r, and set $A_{ij} = r/n$ if i and j are in the same partition and $A_{ij} = 0$ otherwise. Notice that the sum of entries in each row of A is 1, and further, that $\|A\|_F^2 = r$. Indeed, for any partition, $\sum_{ij} A_{ij}^2 = 1$, where the sum is over i, j in this partition. Summing over all partitions, we obtain that $\|A\|_F^2 = r$.

In this case, $f_\theta(\mathbf{1}) = (\theta, \ldots, \theta)$, hence we have

$$\begin{aligned}\psi(\theta_1\mathbf{1}, \beta_1; \theta_0\mathbf{1}, \beta_0) &= \|A\|_F^2 + \left\|(1+a)\mathbf{1} - \mathbf{1} + (-1)A\mathbf{tanh}\left(\frac{1/2}{-1/2 - 1/2}(-a)\mathbf{1} + \mathbf{1}\right)\right\|_2^2 \\ &= \|A\|_F^2 + \|a\mathbf{1} - A\mathbf{tanh}((a/2+1)\mathbf{1})\|_2^2 = \|A\|_F^2 + \|a\mathbf{1} - \mathbf{tanh}((a/2+1)\mathbf{1})\|_2^2 = \|A\|_F^2,\end{aligned}$$

using the fact that the sum of elements in each row of A equals 1, and the fact that $a = \tanh(a/2 + 1)$.

Fix a constant $c_0 > 0$, let $\zeta = c_0/\|A\|_F$, and let $\theta_\zeta = 1 + \zeta a$. By Lemma 19, we have

$$\psi(\theta_\zeta\mathbf{1}, \beta_\zeta; \theta_0\mathbf{1}, \beta_0) = \zeta^2\psi(\theta_1\mathbf{1}, \beta_1; \theta_0\mathbf{1}, \beta_0) = \frac{c_0^2}{\|A\|_F^2}\psi(\theta_1\mathbf{1}, \beta_1; \theta_0\mathbf{1}, \beta_0) = c_0^2.$$

By Lemma 18 one has

$$D_{KL}(\mathbb{P}_{\theta_\zeta, \beta_\zeta}\|\mathbb{P}_{\theta_0, \beta_0}) \leq C\psi(\theta_\zeta\mathbf{1}, \beta_\zeta; \theta_0\mathbf{1}, \beta_0) \leq Cc_0^2.$$

By Pinsker's inequality,

$$D_{TV}(\mathbb{P}_{\theta_\zeta, \beta_\zeta}\|\mathbb{P}_{\theta_0, \beta_0}) \leq \sqrt{D_{KL}(\mathbb{P}_{\theta_\zeta, \beta_\zeta}\|\mathbb{P}_{\theta_0, \beta_0})/2} \leq \sqrt{Cc_0^2/2}.$$

If c_0 is sufficiently small, then by Le-Cam's inequality (Lemma 20), we have that either

$$\mathbb{P}_{h_\zeta,\beta_\zeta}[\|\hat{h} - h_\zeta\| \geq \|h_\zeta - h_0\|/2] \geq 0.49$$

or

$$\mathbb{P}_{h_0,\beta_0}[\|\hat{h} - h_0\| \geq \|h_\zeta - h_0\|/2] \geq 0.49.$$

Further, notice that $\|h_\zeta - h_0\|^2/n = (a\zeta)^2 = a^2 c_0^2/\|A\|_F^2$, as required, which concludes the proof.

D Applications

In this Section, we describe various applications of Theorem 6 in machine learning problems. As we shall see, in order to get the estimation rates in each application, we will often use the weaker bound with respect to $\|A\|_F$. The challenge is thus to compute the covering numbers in each specific case.

D.1 Linear class

In this section, we consider the general setting of logistic regression from dependent observations, first studied in [DDP19]. This model has already been described in Setting 1. We are given a feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and the goal if to find estimation rates for the parameter vector $\theta \in \mathbb{R}^d$ and $\beta \in \mathbb{R}$, using MPLE. We start with Theorem 1. Clearly, the only obstacle to applying Theorem 6 is to calculate the covering numbers. This is exactly what will be done in the proof. The following is just a restatement of Theorem 1.

Theorem 11 (Generalization of [DDP19])**.** *Let σ be sampled according to Setting 1 and suppose $\hat{\theta} \in \mathbb{R}^d, \hat{\beta} \in \mathbb{R}$ are the MPLE estimates for this problem. Then, with probability at least $1-\delta$*

$$\|\hat{\theta} - \theta^*\|_2^2 + |\hat{\beta} - \beta^*|^2 \lesssim \frac{d\log n + \log(1/\delta)}{\|A\|_F^2}.$$

Proof. We will prove the bound for θ, the proof for β is completely analogous. First of all, we notice that to use the result of the general Theorem 6, we need to specify what f_θ is in this case. By the way Setting 1 is defined, is is clear that $f_\theta(x) = x^\top\theta$. Hence, we have that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (f_{\hat{\theta}}(x_i) - f_{\theta^*}(x_i))^2 &= \frac{1}{n}\sum_{i=1}^n (x_i^\top(\hat{\theta} - \theta^*))^2 \\ &= \frac{1}{n}\|\mathbf{X}\theta\|^2 \\ &= \frac{(\hat{\theta} - \theta^*)^\top \mathbf{X}^\top \mathbf{X}(\hat{\theta} - \theta^*)}{n}\end{aligned}$$

By the lower bound κ on the eigenvalues of $\mathbb{X}^\top\mathbf{X}/n$, this implies that

$$\|\hat{\theta} - \theta^*\|^2 \leq \frac{1}{n\kappa}\sum_{i=1}^n (f_{\hat{\theta}}(x_i) - f_{\theta^*}(x_i))^2$$

Thus, by applying Theorem 6, we immediately obtain that with probability $\geq 1-\delta$

$$\|\hat{\theta} - \theta^*\|^2 \lesssim \frac{\inf_{\epsilon\geq 0}\left(\log\frac{n}{\delta} + \epsilon n + \log N(\mathcal{F}, X, \epsilon)\right)}{\|A\|_F^2}$$

To conclude the proof, we need to bound $N(\mathcal{F}, X, \epsilon)$ and choose a suitable ϵ. As we know, $\mathcal{F}$ is a set of functions f_θ indexed by θ, together with a metric d. Since for all $\|x_i\| \leq M$, we have

$$d(f_{\theta_1}, f_{\theta_2}) = \sqrt{\frac{\sum_{i=1}^n (x_i^\top(\theta_1 - \theta_2))^2}{n}} \leq M\|\theta_1 - \theta_2\|$$

Hence, finding an ϵ-net for $\mathcal{F}$ with respect to d is equivalent to finding an ϵ/M-net of the space of θ's with respect to the l_2-norm. By the assumptions of Setting 1, we know that θ lies in a ball of radius 1. Hence, the size of an ϵ-net for the unit ball is known([Ver18]) to be at most $(1/\epsilon)^d$, since the space of θ's lies in $\mathbb{R}^d$. To conclude, we have

$$\log N(\mathcal{F}, X, \epsilon) \lesssim d \log \frac{1}{\epsilon}$$

Setting $\epsilon = 1/n$, we have that

$$\log \frac{n}{\delta} + \epsilon n + \log N(\mathcal{F}, X, \epsilon) \lesssim d \log n + \log \frac{1}{\delta},$$

from which the rate for θ follows. For β, we can use the same approach, first using the bound of Theorem 6 and then bounding the covering number in exactly the same way. □

However, improving the rate of [DDP19] is just one of the applications of the general Theorem for logistic regression. After all, the assumption $\|A\|_F = \Omega(\sqrt{n})$ seems too restrictive when we care about estimating θ. If anything, having weaker interactions among the spins should help us recover θ more easily, since we are closer to the vanilla logistic regression setting. In Theorem 4, we show that if we pick the matrix $\mathbf{X}$ randomly, sampling each coordinate i.i.d. from the Gaussian distribution, then with high probability we get a $1/\sqrt{n}$ rate of consistency, regardless of the matrix A. To do this, we will use the general guarantee provided by Theorem 6. First, we state and prove a Lemma that brings this general rate to a more convenient form.

Lemma 21. *Let $\mathcal{C}_1'$ be the quantity defined in* (16)*. Then,*

$$\mathcal{C}_1'(\mathcal{H}, h^*, \beta^*) \leq \frac{1}{n \cdot \inf_{\lambda \in \mathbb{R}, h \in \mathcal{H}} \left(\|\lambda A\|_F^2 + \left\| \frac{h^* - h}{\|h^* - h\|} - \lambda A \tanh\left(\frac{\beta^*}{\lambda} \frac{h^* - h}{\|h^* - h\|} + h^* \right) \right\|^2 \right)}.$$

Proof. Let $\beta \in [-M, M]$ and $h \in \mathcal{H}$. Then,

$$\begin{aligned}
\frac{\|h - h^*\|_2^2}{\psi(h, \beta)} &= \frac{\|h - h^*\|_2^2}{(\beta - \beta^*)^2 \|A\|_F^2 + \left\| h^* - \hat{h} + (\beta^* - \hat{\beta}) A \tanh\left(\frac{\beta^*}{\hat{\beta} - \beta^*} (h^* - \hat{h}) + h^* \right) \right\|_2^2} \\
&= \frac{1}{\left(\frac{\hat{\beta} - \beta^*}{\|\hat{h} - h^*\|} \right)^2 \|A\|_F^2 + \left\| \frac{h^* - \hat{h}}{\|h^* - \hat{h}\|} + \frac{\beta^* - \hat{\beta}}{\|h^* - \hat{h}\|} A \tanh\left(\frac{\beta^*}{\frac{\hat{\beta} - \beta^*}{\|h^* - \hat{h}\|}} \frac{h^* - \hat{h}}{\|h^* - \hat{h}\|} + h^* \right) \right\|^2} \\
&= \frac{1}{\inf_{\lambda \in \mathbb{R}} \|\lambda A\|_F^2 + \left\| \frac{h^* - \hat{h}}{\|h^* - \hat{h}\|} + \lambda A \tanh\left(\frac{\beta^*}{\lambda} \frac{h^* - \hat{h}}{\|h^* - \hat{h}\|} + h^* \right) \right\|^2}.
\end{aligned} \tag{39}$$

Dividing both sides of the inequality by n and taking supremum over $\beta \in [-M, M]$ and $h \in \mathcal{H}$, the result follows. □

Now, we are ready to prove Theorem 4. We present the statement in a slightly more detailed way below as Theorem 12.

Theorem 12. *Suppose σ is sampled according to Setting 1. We drop all assumptions about $\mathbf{X}$ and instead assume that we sample $\mathbf{X}_{ij} \sim N(0, 1)$ independently for all $i \in [n], j \in [d]$. Let $\hat{\theta}, \hat{\beta}$ be the estimates of the MPLE. Then, with probability $1 - \delta$*

$$\|\hat{\theta} - \theta^*\| \leq C(M) \mathrm{e}^{\sqrt{\log n}} \frac{\sqrt{d \log n + \log(1/\delta)}}{\sqrt{n}} \max\left(\frac{\|A\|_2 \sqrt{\log(1/\delta) + \log n + d \log\log n + d \log(\log(1/\delta))}}{\|A\|_F}, 1 \right)$$

Consequently, $\hat{\theta}$ is an α_n-consistent estimator, where

$$\alpha_n = \mathrm{e}^{\sqrt{\log n}} \sqrt{\frac{d \log n}{n}} \max\left(\frac{\sqrt{d \log\log n + \log n} \|A\|_2}{\|A\|_F}, 1 \right)$$

Proof. We use the rate for θ provided in Theorem 6, combined with Lemma 21, which simplifies the rate. This gives us that with probability $\geq 1-\delta$:

$$\|\mathbf{X}\hat{\theta} - \mathbf{X}\theta^*\|^2 \lesssim \frac{\log\frac{n}{\delta} + \inf_{\epsilon\geq 0}\left(n\epsilon + \log\ N\left(\mathcal{F}, X, \epsilon\right)\right)}{\inf_{\lambda\in\mathbb{R},\|\theta\|\leq M}\left(\|\lambda A\|_F^2 + \left\|\frac{\mathbf{X}\theta^*-\mathbf{X}\theta}{\|\mathbf{X}\theta^*-\mathbf{X}\theta\|} - \lambda A\tanh\left(\frac{\beta^*}{\lambda}\frac{\mathbf{X}\theta^*-\mathbf{X}\theta}{\|\mathbf{X}\theta^*-\mathbf{X}\theta\|} + \mathbf{X}\theta^*\right)\right\|^2\right)} \tag{40}$$

We will first show that the denominator is bounded with high probability. For convenience, since we want a bound for $\|\mathbf{X}\hat{\theta} - \mathbf{X}\theta^*\|$, we will work with the denominator without the squares. This is without loss of generality, since for $a, b > 0$ we have $a^2 + b^2 = \Theta((a+b)^2)$. First, a simple scale invariance of the expression yields

$$\begin{aligned}
&\inf_{\lambda\in\mathbb{R},\|\theta\|\leq 1}\left(\|\lambda A\|_F + \left\|\frac{\mathbf{X}\theta^*-\mathbf{X}\theta}{\|\mathbf{X}\theta^*-\mathbf{X}\theta\|} - \lambda A\tanh\left(\frac{\beta^*}{\lambda}\frac{\mathbf{X}\theta^*-\mathbf{X}\theta}{\|\mathbf{X}\theta^*-\mathbf{X}\theta\|} + \mathbf{X}\theta^*\right)\right\|\right)\\
&\quad= \frac{1}{\|\mathbf{X}\theta^*-\mathbf{X}\theta\|}\inf_{\lambda\in\mathbb{R},\|\theta\|\leq 1}\left(\|\lambda A\|_F + \left\|\mathbf{X}\theta^*-\mathbf{X}\theta - \lambda A\tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^*-\mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\|\right)
\end{aligned} \tag{41}$$

Let's focus on the infimum of the latter expression. In particular, we fix θ with bounded norm. Also, notice that we can replace A by $A/\|A\|_2$ and β^* by $\beta^*\|A\|$ and the probability model remains the same. This is done without loss of generality to ensure that A has unit spectral norm, which will simplify the computations later on. Notice that since $\|A\|$ is bounded, β^* remains bounded after the transformation. Also, notice that after this transformation $\|A\|_F = \Omega(1)$.

We will show that for all values of λ, at least one of the two terms is large. First of all, by triangle inequality

$$\left\|\mathbf{X}\theta^*-\mathbf{X}\theta - \lambda A\tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^*-\mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\| \geq \|\mathbf{X}\theta^*-\mathbf{X}\theta\| - \left\|\lambda A\tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^*-\mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\|$$

We will show that with high probability over the choice of $\mathbf{X}$, this difference is large. Let's consider $\|\mathbf{X}(\theta^*-\theta)\|$ first. Notice that this each coordinate of the vector $\mathbf{X}(\theta^*-\theta)$ is independent from the rest and distributed as a gaussian $N(0, \|\theta-\theta^*\|^2)$. Thus, we can apply Theorem 3.1.1 of [Ver18], which gives us that

$$\left\|\|\mathbf{X}(\theta^*-\theta)\| - \|\theta-\theta^*\|\sqrt{n}\right\|_{\psi_2} \leq K\|\theta-\theta^*\| \tag{42}$$

for some constant K. This means that the norm of the vector is well concentrated around $\|\theta-\theta^*\|\sqrt{n}$. Now, let's consider the second term. For simplicity, denote by $\mathbf{u} = \frac{\beta^*}{\lambda}(\mathbf{X}\theta^*-\mathbf{X}\theta) + \mathbf{X}\theta^*$. In the computations below, $\mathbb{E}$ refers to expectation with respect to the distribution of $\mathbf{X}$.

$$\begin{aligned}
\mathbb{E}\left\|A\tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^*-\mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\|^2 &= \mathbb{E}\sum_i\left(\sum_j A_{ij}\tanh(\mathbf{u}_j)\right)^2\\
&= \mathbb{E}\sum_i\sum_j A_{ij}^2\tanh(\mathbf{u}_j)^2 \leq \|A\|_F^2
\end{aligned}$$

We used the fact that the coordinates of $\mathbf{u}$ are all independent from each other and have mean 0. This means that the coordinates of $\tanh(\mathbf{u})$ are also independent and have mean 0. We conclude that

$$\mathbb{E}\left\|A\tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^*-\mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\| \leq \sqrt{\mathbb{E}\left\|A\tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^*-\mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\|^2} \leq \|A\|_F$$

We just bounded the mean of the function $f(\mathbf{X}) = \left\|A\tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^*-\mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\|$. The next step is to show that it is concentrated around that value. To do that, we will use the gaussian isoperimetric inequality,

see e.g. Theorem 5.1 in [BLM13]. We can view $\mathbf{X}$ as a $n \times d$ vector of independent normal gaussian variables. The 2 norm of this vector is the frobenius norm of $\mathbf{X}$. Hence, to use the isoperimetric inequality, we need to bound the Lipschitzness of f w.r.t. the frobenius norm of $\mathbf{X}$. Let $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^{n\times d}$ be two matrices. We denote by $\mathbf{X}_i$ as the i-th row of matrix $\mathbf{X}$. Then, the difference in value is

$$\begin{aligned}
|f(\mathbf{X}) - f(\mathbf{X}')| &\leq \left\| A \tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^* - \mathbf{X}\theta) + \mathbf{X}\theta^*\right) - A\tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}'\theta^* - \mathbf{X}'\theta) + \mathbf{X}'\theta^*\right)\right\| \\
&\leq \|A\| \left\| \tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^* - \mathbf{X}\theta) + \mathbf{X}\theta^*\right) - \tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}'\theta^* - \mathbf{X}'\theta) + \mathbf{X}'\theta^*\right)\right\| \\
&\leq \|A\| \left\| \frac{\beta^*}{\lambda}(\mathbf{X}\theta^* - \mathbf{X}\theta) + \mathbf{X}\theta^* - \frac{\beta^*}{\lambda}(\mathbf{X}'\theta^* - \mathbf{X}'\theta) + \mathbf{X}'\theta^*\right\| \\
&\leq \|A\| \left(\frac{\beta^*}{|\lambda|} \sqrt{\sum_{i=1}^n ((\mathbf{X}_i - \mathbf{X}'_i)^\top(\theta^* - \theta))^2} + \sqrt{\sum_{i=1}^n ((\mathbf{X}_i - \mathbf{X}'_i)^\top \theta^*)^2}\right) \\
&\leq \|A\| \left(\frac{\beta^*}{|\lambda|} \|\theta^* - \theta\| \|\mathbf{X} - \mathbf{X}'\|_F + \|\theta^*\| \|\mathbf{X} - \mathbf{X}'\|_F \right) \\
&= \|A\| \left(\frac{\beta^*}{|\lambda|} \|\theta^* - \theta\| + \|\theta^*\| \right) \|\mathbf{X} - \mathbf{X}'\|_F
\end{aligned}$$

Now, since the entries of $\mathbf{X}$ are i.i.d. gaussians and $\|A\|, \|\theta^*\|$ are bounded by constants, from the gaussian isoperimetric inequality ([Ver18; BLM13]) we get that with probability $1 - \mathrm{e}^{-t^2/2}$

$$\begin{aligned}
\left\| \lambda A \tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^* - \mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\| &\leq \mathbb{E}\left\| \lambda A \tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^* - \mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\| + ct(\|\theta^* - \theta\| + \lambda) \\
&\leq \|\lambda A\|_F + ct(\|\theta^* - \theta\| + \lambda)
\end{aligned}$$

Also, by (42) we have that with probability $1 - \mathrm{e}^{-t^2/2}$

$$\|\mathbf{X}\theta^* - \mathbf{X}\theta\| \geq \|\theta^* - \theta\|\sqrt{n} - tK\|\theta - \theta^*\|$$

Hence, we get that with probability $1 - \mathrm{e}^{-t^2}$

$$\left\| \mathbf{X}\theta^* - \mathbf{X}\theta - \lambda A \tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^* - \mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\| \geq c_1\|\theta^* - \theta\|\sqrt{n} - \|\lambda A\|_F - tC\|\theta - \theta^*\| - tc\lambda$$

Now we distinguish cases for λ. First of all, suppose that

$$\lambda \leq C\sqrt{n}\|\theta - \theta^*\| \min\left(\frac{1}{\sqrt{\log(1/\delta') + \frac{\log n}{2}}}, \frac{1}{\|A\|_F}\right) := \kappa$$

for a suitable constant C. Then, applying the previous result, with probability at least $1 - (\delta'/\sqrt{n})$ we have that

$$\left\| \mathbf{X}\theta^* - \mathbf{X}\theta - \lambda A \tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^* - \mathbf{X}\theta) + \mathbf{X}\theta^*\right)\right\| \geq C'\|\theta^* - \theta\|\sqrt{n} \tag{43}$$

On the other hand, if

$$\lambda \geq \kappa$$

then the first term of the expression is

$$\|\lambda A\|_F \geq C\sqrt{n}\|\theta - \theta^*\| \min\left(\frac{\|A\|_F}{\sqrt{\log(1/\delta') + \frac{\log n}{2}}}, 1\right)$$

Note that this bound holds with probability 1. Now denote

$$g(\lambda) = \left\| \mathbf{X}\theta^* - \mathbf{X}\theta - \lambda A \tanh\left(\frac{\beta^*}{\lambda}(\mathbf{X}\theta^* - \mathbf{X}\theta) + \mathbf{X}\theta^* \right) \right\|$$

To get a bound for the infimum, we need the bound of (43) to hold for all $|\lambda| \leq \kappa$. To do that, we notice that the expression of the infimum is Lipschitz with respect to λ and the Lipschitz constant is at most $C''\sqrt{n}$ for some constant C'' that depends on $\|A\|_F, \|A\|_2$. Now suppose we pick an ϵ-net for the set of $|\lambda| \leq \kappa$. Clearly, we can pick such a net that has at most $2\kappa/\epsilon$ elements. Suppose that for a point in the net λ_1 we have $g(\lambda_1) \geq C'\|\theta^* - \theta\|\sqrt{n}$ and consider another point λ_2 with $|\lambda_2 - \lambda_1| < \epsilon$. We then have

$$f(\lambda_2) \geq f(\lambda_1) - C''\sqrt{n}\epsilon \geq C'\|\theta^* - \theta\|\sqrt{n} - C''\sqrt{n}\epsilon$$

Hence, in order for to obtain a similar bound for $f(\lambda_2)$, we need to choose $\epsilon = \Theta(\|\theta^* - \theta\|)$. The size of this net will be $O(\kappa/\epsilon) = O(\sqrt{n})$. Hence, if (43) holds for all points in this net, it will hold for all $|\lambda| \leq \kappa$. The probability that (43) holds for all points in the net is at least $1 - \delta'$, which easily follows from a union bound. Hence, with probability at least $1 - \delta'$ we have that

$$\begin{aligned} &\inf_{\lambda \in \mathbb{R}} \left(\|\lambda A\|_F + \left\| \frac{\mathbf{X}\theta^* - \mathbf{X}\theta}{\|\mathbf{X}\theta^* - \mathbf{X}\theta\|} - \lambda A \tanh\left(\frac{\beta^*}{\lambda} \frac{\mathbf{X}\theta^* - \mathbf{X}\theta}{\|\mathbf{X}\theta^* - \mathbf{X}\theta\|} + \mathbf{X}\theta^* \right) \right\| \right) \\ &\qquad = \Omega\left(\sqrt{n}\|\theta - \theta^*\| \min\left(\frac{\|A\|_F}{\sqrt{\log(1/\delta') + \log n}}, 1 \right) \right) \end{aligned} \tag{44}$$

Now, from (41) we also need to upper bound $\|\mathbf{X}\theta - \mathbf{X}\theta^*\|$. By the concentration of the norm, which we have already used, we get that with probability $1 - \mathrm{e}^{-t^2}$

$$\|\mathbf{X}\theta^* - \mathbf{X}\theta\| \leq \|\theta^* - \theta\|\sqrt{n} + tK\|\theta - \theta^*\|$$

Hence, with probability $1 - \delta'$

$$\|\mathbf{X}\theta^* - \mathbf{X}\theta\| \leq C'''\|\theta^* - \theta\|(\sqrt{n} + \sqrt{\log(1/\delta')})$$

Combining this with (44), we get that for any $\delta' \geq \mathrm{e}^{-n}$, with probability $1 - \delta'$ we have

$$\begin{aligned} &\inf_{\lambda \in \mathbb{R}} \left(\|\lambda A\|_F + \left\| \frac{\mathbf{X}\theta^* - \mathbf{X}\theta}{\|\mathbf{X}\theta^* - \mathbf{X}\theta\|} - \lambda A \tanh\left(\frac{\beta^*}{\lambda} \frac{\mathbf{X}\theta^* - \mathbf{X}\theta}{\|\mathbf{X}\theta^* - \mathbf{X}\theta\|} + \mathbf{X}\theta^* \right) \right\| \right) \\ &\qquad = \Omega\left(\min\left(\frac{\|A\|_F}{\sqrt{\log(1/\delta') + \log n}}, 1 \right) \right) \end{aligned}$$

Now, to conclude the analysis of the denominator of the rate, we need to take the infimum over all θ with bounded norm. First of all, consider the set

$$\mathcal{U} = \left\{ \frac{\mathbf{X}\theta^* - \mathbf{X}\theta}{\|\mathbf{X}\theta^* - \mathbf{X}\theta\|} : \|\theta\| \leq 1 \right\}$$

We will follow a covering argument for $\mathcal{U}$ similar to what we did previously. First of all, we prove that for $\mathbf{a}, \mathbf{a}' \in \mathcal{U}$, if these two vectors are close, then the value of the infimum does not change much. Notice that to prove that, it is enough to prove that for a fixed λ, the expression doesn't change much. For a fixed λ, we have

$$\left| \left\| \mathbf{a} - \lambda A \tanh\left(\frac{\beta^*}{\lambda}\mathbf{a} + \mathbf{X}\theta^* \right) \right\| - \left\| \mathbf{a}' - \lambda A \tanh\left(\frac{\beta^*}{\lambda}\mathbf{a}' + \mathbf{X}\theta^* \right) \right\| \right|$$

$$\leq \|\mathbf{a} - \mathbf{a}'\| + |\lambda|\|A\| \left\| \tanh\left(\frac{\beta^*}{\lambda}\mathbf{a} + \mathbf{X}\theta^*\right) - \tanh\left(\frac{\beta^*}{\lambda}\mathbf{a}' + \mathbf{X}\theta^*\right)\right\|$$
$$\leq (1 + \beta^*\|A\|)\|\mathbf{a} - \mathbf{a}'\|$$

Thus, the infimum is a Lipschitz function of $\mathbf{a}$ and the Lipschitzness is a constant. Now let's denote

$$\gamma = \min\left(\frac{\|A\|_F}{\sqrt{\log(1/\delta') + \log n}}, 1\right)$$

Consider an ϵ-net on the set $\mathcal{U}$ with respect to the 2-norm. If the value of the infimum is at least γ for all the points in the net, then any other point has value at least $\gamma - (1 + \beta^*\|A\|)\epsilon$. Thus, if we choose $\epsilon = \Theta(\gamma)$, then we get that for all $\mathbf{a} \in \mathcal{U}$, the infimum is $\Omega(\gamma)$. To determine the probability of this event, we can simply take a union bound over all $\mathcal{N}(\mathcal{U}, \|\cdot\|_2, \gamma)$ points in the net. Thus, our task reduces to finding a bound for this covering number. To do that, we need a few facts about the singular values of the matrix $\mathbf{X}$. We use a result from [Ver10]. According to this, with probability at least $1 - \mathrm{e}^{-ct^2}$ we have

$$\left\|\frac{1}{n}\mathbf{X}^\top\mathbf{X} - I\right\| \leq \sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}$$

From now on, we will assume this is the case without explicitly mentioning the probability. This implies that with probability at least $1 - \mathrm{e}^{-cn}$, the eigenvalues of $\mathbf{X}^\top\mathbf{X}/n$ are upper and lower bounded by constants $\lambda_{min}, \lambda_{max}$.

By definition of the set $\mathcal{U}$, it is enough to assume without loss of generality that θ with $\|\theta^* - \theta\| = 1$ for all θ of interest(simply divide by the norm $\|\theta^* - \theta\|$. Now, consider θ_1, θ_2 with this property. We have that

$$\begin{aligned}\left|\frac{\mathbf{X}\theta^* - \mathbf{X}\theta_1}{\|\mathbf{X}\theta^* - \mathbf{X}\theta_1\|} - \frac{\mathbf{X}\theta^* - \mathbf{X}\theta_2}{\|\mathbf{X}\theta^* - \mathbf{X}\theta_2\|}\right| &\leq \frac{\|\mathbf{X}\theta_2 - \mathbf{X}\theta_1\|}{\min(\|\mathbf{X}\theta^* - \mathbf{X}\theta_1\|, \|\mathbf{X}\theta^* - \mathbf{X}\theta_2\|)} \\ &\leq \frac{\sqrt{n\lambda_{max}}\|\theta_2 - \theta_1\|}{\sqrt{n\lambda_{min}}} \\ &= O(\|\theta_2 - \theta_1\|)\end{aligned}$$

Thus, to get an ϵ-net for $\mathcal{U}$, it suffices to construct an $O(\epsilon)$-net for $\theta^* + B(0,1)$, where $B(0,1)$ is the unit sphere in $\mathbb{R}^d$. However, it is known that for this set, we can construct an ϵ-net with $O((1/\epsilon)^d)$ elements. From our previous analysis, we should choose $\epsilon = O(\gamma)$. Thus, by applying a union bound over the points in the net and then using the Lipschitzness, we get that with probability at least $1 - (1/\gamma)^d\delta'$,

$$\inf_{\lambda\in\mathbb{R}, \|\theta\|=O(1)} \left(\|\lambda A\|_F + \left\|\frac{\mathbf{X}\theta^* - \mathbf{X}\theta}{\|\mathbf{X}\theta^* - \mathbf{X}\theta\|} - \lambda A \tanh\left(\frac{\beta^*}{\lambda}\frac{\mathbf{X}\theta^* - \mathbf{X}\theta}{\|\mathbf{X}\theta^* - \mathbf{X}\theta\|} + \mathbf{X}\theta^*\right)\right\|\right) = \Omega(\gamma)$$

We would like to choose δ' so that the probability of error is at most δ. This gives us

$$(1/\gamma)^d\delta' < \delta \implies d\log(\gamma) + \log(1/\delta') > \log(1/\delta)$$

If $\gamma = 1$, this inequality reduces to $\delta' < \delta$. Otherwise, since $\|A\|_F \geq 1$, it suffices to satisfy the inequality

$$\log(1/\delta') - \frac{d}{2}\log(\log(1/\delta') + \log n) > \log(1/\delta)$$

We notice that the preceding inequality is satisfied if we set

$$\log(1/\delta') = \log(1/\delta) + d\log(\log(1/\delta) + \log n)$$

Hence, we conclude that with probability at least $1 - \delta$, the infimum is at least

$$\min\left(\frac{\|A\|_F}{\sqrt{\log(1/\delta) + \log n + d\log(\log(1/\delta) + \log n)}}, 1\right)$$

$$= \Omega\left(\min\left(\frac{\|A\|_F}{\sqrt{\log(1/\delta) + \log n + d\log\log n + d\log(\log(1/\delta))}}, 1\right)\right)$$

Now, all that remains is to compute the numerator of the rate. Notice that we are in the event where the eigenvalues of $\mathbb{X}^\top\mathbf{X}/n$ are bounded by positive constants. Hence, suppose M is an upper bound for the eigenvalues. We have

$$\begin{aligned} d(f_{\theta_1}, f_{\theta_2}) &= \sqrt{\frac{\|\mathbf{X}\theta\|^2}{n}} \\ &\leq M\|\theta_2 - \theta_1\| \end{aligned}$$

This means that it is enough to find an ϵ-net for the space of θ's with respect to the l_2-norm. This computation was also done in Theorem 11 and the value we found was

$$\sqrt{d\log n + \log(1/\delta)}$$

if we want error probability δ.

Hence, we get that for any $\delta > \mathrm{e}^{-n}$, with probability $1-\delta$ we have:

$$\|\mathbf{X}\hat{\theta} - \mathbf{X}\theta^*\| \lesssim \sqrt{d\log n + \log(1/\delta)}\max\left(\frac{\sqrt{\log(1/\delta) + \log n + d\log\log n + d\log(\log(1/\delta))}}{\|A\|_F}, 1\right)$$

Now, we need to consider the constant $C(M)$ that is hidden with $\lesssim$. The result of Theorem 6 states that $C(M)$ depends single-exponentially on the bound M. In our case, the parameters are bounded. However, $\mathbf{X}\theta$ is a random quantity, hence it's maximum entry is not necessarily bounded. We address this issue now. We immediately notice that each coordinate of $\mathbf{X}\theta$ is a gaussian with 0 mean and variance at most M. Also, the coordinate values are all independent, since the rows of $\mathbf{X}$ are all independent. Hence, the maximum coordinate is distributed as the maximum of n independent gaussian variables, so it's expectation is at most $M\sqrt{2\log n}$. Also, we know that the maximum is also well concentrated, which means that with probability at least $1-\delta$,

$$\|\mathbf{X}\theta\|_\infty \leq M\sqrt{2\log n} + \sqrt{\log(1/\delta)}$$

Now, if $\delta > 1/\mathrm{poly}(n)$, we have that

$$\|\mathbf{X}\theta\|_\infty = O(\sqrt{\log n})$$

This implies that the constant in the rate is at most $O\left(\mathrm{e}^{\sqrt{\log n}}\right)$, which is smaller than any polynomial of n. The result now follows by recalling that we normalized A by $\|A\|_2$ in the beginning.

□

Remark 1. Since $\|A\|_F \geq \|A\|_2$, Corollary 12 always guarantees a $\tilde{O}(d/\sqrt{n})$ consistency rate, regardless of the norms of the matrix A. We have ignored the $\exp(\sqrt{\log n})$ part of the rate, since it is smaller than any polynomial of n.

D.2 Sparse Logistic Regression

Another interesting application of the general Theorem 6 involves the case of Sparse Logistic Regression. In this case, we have the familiar Setting 1 of logistic regression, with an unknown vector $\theta \in \mathbb{R}^d$ and a fixed matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$. The extra assumption comes in the form of sparsity of the vector θ. We will assume, as is standard in the literature, that θ is l_1-sparse. This also enables us to run MPLE efficiently. Clearly, we expect the rate to depend on s in that case. The quantity that determines the rate is the metric entropy for sparse subsets of $\mathbb{R}^d$. For convenience, we restate it to include a more detailed description of the assumptions.

Theorem 13. *Let σ be sampled according to setting 1. Additionally, assume that $\|\theta^*\|_1 \leq s$. We also assume the restricted eigenvalue property that is stated in Setting 1, which means that for all $\|\theta\|_1 \leq s$ we have*

$$\|\mathbf{X}\theta\| \geq \kappa\sqrt{n}\|\theta\|$$

where κ is a constant. Let $\hat{\theta}, \hat{\beta}$ be the estimates we get for MPLE constrained in the sparse region. Then, w.pr. $\geq 1-\delta$,

$$\|\hat{\theta} - \theta^*\|_2^2 + |\hat{\beta} - \beta^*|^2 \lesssim \frac{(n^2 s \log(d))^{1/3} + \log(1/\delta)}{\|A\|_F^2}.$$

Proof. Notice that since $\hat{\theta}, \theta^*$ are sparse, by the restricted eigenvalue property we get

$$\begin{aligned}\frac{\sum_{i=1}^n (f_{\hat{\theta}} - f_{\theta^*})^2}{n} &= \frac{\|\mathbf{X}(\hat{\theta} - \theta^*)\|^2}{n} \\ &\geq \kappa^2 \|\hat{\theta} - \theta^*\|\end{aligned}$$

Thus, we can immediately apply Theorem 6 to obtain with probability $1-\delta$

$$\|\hat{\theta} - \theta^*\|_2^2 + |\hat{\beta} - \beta^*|^2 \lesssim \frac{\inf_{\epsilon \geq 0} \left(\log \frac{n}{\delta} + \epsilon n + \log N(\mathcal{F}, X, \epsilon) \right)}{\|A\|_F^2}$$

All that remains now is to bound the covering number $N(\mathcal{F}, X, \epsilon)$. Since we are in Setting 1, the norms $\|x_i\|$ are uniformly bounded. This means that, similarly to Theorems 11, 12, we have that

$$d(f_{\theta_1}, f_{\theta_2}) \leq M\|\theta_1 - \theta_2\|$$

Thus, is suffices to compute the covering number of the space of θ's with respect to the l_2-norm. The space of θ's is $\mathcal{U} = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq s\}$. Thus, we need the covering numbers for this set. A result from [RWY11] states that

$$\log N(\mathcal{U}, \|\cdot\|_2, \epsilon) \leq C \frac{s \log d}{\epsilon^2}$$

Thus, to find the optimal ϵ, we should minimize the expression

$$\epsilon n + \frac{s \log d}{\epsilon^2}.$$

This is clearly minimized when the two terms are equal, which means

$$\epsilon n = \frac{s \log d}{\epsilon^2} \implies \epsilon = \left(\frac{s \log d}{n} \right)^{1/3}$$

Hence, the optimal value for the numerator is

$$\log(1/\delta) + (n^2 s \log(d))^{1/3}$$

The result follows. □

D.3 Neural Network Regression

Another application of Theorem 6 involves the case where the regression function f_θ is not linear, but a neural network. Since neural networks are often overparametrized, we cannot hope to recover the parameter θ is this case. However, we can still apply Theorem 6, which will yield guarantees for estimating the output of the network. We assume we are in Setting 2, which means that all the parameters of the network are bounded. Again, the crucial quantity is the metric entropy for the image of these neural networks, for which we have sufficiently strong guarantees from the literature.

Theorem 5. *Suppose Setting 2, and let* $K^2 = \frac{1}{n}\sum_i \|x_i\|_2^2$. *Then, with probability* $\geq 1-\delta$,

$$\frac{1}{n}\sum_{i=1}^{n}(f_{\hat{\theta}}(x_i) - f_{\theta^*}(x_i))^2 + |\hat{\beta} - \beta^*|^2 \quad \lesssim \frac{(n^2K^2R^2\log d)^{1/3} + \log\left(\frac{n}{\delta}\right)}{\|A\|_F^2}.$$

Proof. To begin with, we can directly apply Theorem 6 to obtain with probability $\geq 1-\delta$:

$$\frac{\sum_{i=1}^{n}(f_{\hat{\theta}} - f_{\theta^*})^2}{n} + |\hat{\beta} - \beta^*|^2 \lesssim \frac{\inf_{\epsilon\geq 0}\left(\log\frac{n}{\delta} + \epsilon n + \log N(\mathcal{F}, X, \epsilon)\right)}{\|A\|_F^2}$$

Hence, we only have to bound the covering numbers of the set $\mathcal{F}$. By the way we have defined distance d, it is essentially the l_2-norm of the difference in the outputs of two networks, divided by $\sqrt{n}$. Hence, we define $\mathcal{H}_X$ to be the set of outputs of neural networks satisfying the bounds of Setting 2, when the input is the matrix X, composed of features $x_1, \ldots, x_n$. Then, it suffices to find an $\epsilon\sqrt{n}$-net for $\mathcal{H}_X$ with respect to the l_2-norm. Such a bound is given in Theorem 3.3 of [BFT17], which gives us (using the fact that we have width bounded by d)

$$\log N(\mathcal{F}, X, \epsilon) \leq \frac{\log d\sum_{i=1}^{n}\|x_i\|^2}{(\sqrt{n}\epsilon)^2}R^2 = \frac{nK^2R^2\log d}{n\epsilon^2}$$

Clearly, to find the ϵ that minimizes the numerator, we set

$$n\epsilon = \frac{K^2R^2\log d}{\epsilon^2} \implies \epsilon = \left(\frac{K^2R^2\log d}{n}\right)^{1/3}$$

Using this value of ϵ, the numerator becomes

$$\log(1/\delta) + (n^2K^2R^2\log d)^{1/3},$$

which concludes the proof. □

D.4 The Curie-Weiss model

The power of obtaining separate rates for estimation of parameters θ, β can be easily showcased when considering the Curie Weiss model with external field. In this model, the external field has a fixed direction $h \in \mathbb{R}^n$. Without loss of generality, we can assume that A is the all $1/n$'s matrix, since the error introduced by this modification is $O(1/n)$. The probability distribution becomes

$$\mathbb{P}[\sigma = y] \propto \exp\left(\theta^*\sum_{i=1}^{n} y_i h_i + \frac{\beta^*}{n}\sum_{i,j} y_i y_j\right) \tag{45}$$

We will identify cases where it is possible to estimate θ or β. For simplicity, we will assume that the coordinates of h lie in $\{-1, 1\}$. The result of [DDP19] does not give us any rate, since $\|A\|_F = 1$. However, as we will see, the property that determines the rate has to do with how much h is close to being an eigenvector of A.

Corollary 1. *Let* σ *be sampled according to* (45) *with* $|\beta^*|, |\theta^*| \leq M$. *Suppose the external field* h *has a fraction* α *of coordinates equal to* 1 *and a fraction* $1-\alpha$ *equal to* -1. *Let* $\hat{\theta}, \hat{\beta}$ *be the estimates when we run MPLE on* $[-M, M]^2$. *Then, with probability at least* $1-\delta$

$$|\hat{\theta} - \theta^*| \leq C\frac{\sqrt{\log n + \log(1/\delta)}}{\sqrt{4n(1-(2\alpha-1)^2)}}$$

Proof. Let $U, V \setminus U$ be the subsets of vertices that have $h_i = 1$ and $h_i = -1$ respectively. Let's start with acquiring a rate for θ^*. Since each parameter is one dimensional and bounded, according to Theorem REF, with probability $1 - \delta$

$$|\hat{\theta} - \theta^*|\|h\| \leq C \frac{\sqrt{\log n + \log(1/\delta)}}{\inf_{\lambda \in \mathbb{R}} \left(|\lambda| + \left\| \frac{h}{\|h\|} - \lambda A \tanh \left(\frac{\beta^*}{\lambda} \frac{h}{\|h\|} + \theta^* h \right) \right\| \right)}$$

which implies that

$$|\hat{\theta} - \theta^*| \leq C \frac{\sqrt{\log n + \log(1/\delta)}}{\inf_{\lambda \in \mathbb{R}} \left(|\lambda| + \left\| h - \lambda A \tanh \left(\frac{\beta^*}{\lambda} h + \theta^* h \right) \right\| \right)}$$

We have that

$$\inf_{\lambda \in \mathbb{R}} \left(|\lambda| + \left\| h - \lambda A \tanh \left(\frac{\beta^*}{\lambda} h + \theta^* h \right) \right\| \right) \geq \inf_{\lambda \in \mathbb{R}} \left\| h - \lambda A \tanh \left(\frac{\beta^*}{\lambda} h + \theta^* h \right) \right\|$$

The tanh vector has a special structure. Specifically, coordinates corresponding to nodes in U have a common value $\mu_1(\lambda)$ and nodes in $V \setminus A$ have a common value $\mu_2(\lambda)$. Hence, if we denote by $\mathbf{1}$ the all ones vector, we have

$$\begin{aligned} \inf_{\lambda \in \mathbb{R}} \left\| h - \lambda A \tanh \left(\frac{\beta^*}{\lambda} h + \theta^* h \right) \right\| &= \inf_{\lambda \in \mathbb{R}} \left(\left\| h - \lambda \left(\frac{\alpha n \mu_1(\lambda) + (1-\alpha) n \mu_2(\lambda)}{n} \mathbf{1} \right) \right\|_2 \right) \\ &\geq \inf_{\lambda \in \mathbb{R}} \left(\|\lambda \mathbf{1} + h\|_2 \right) \end{aligned}$$

The latter quantity is the distance of h from the subspace spanned by $\mathbf{1}$. Hence, the optimal λ^* corresponds to the projection of h onto this subspace. By standard calculations, we have that

$$\lambda^* = -\frac{\langle h, \mathbf{1} \rangle}{\|\mathbf{1}\|^2} = -\frac{2\alpha n - n}{n} = 1 - 2\alpha$$

and

$$\inf_{\lambda \in \mathbb{R}} \left(\|\lambda \mathbf{1} + h\|_2 \right) = \sqrt{\|h\|_2^2 - \frac{\langle h, \mathbf{1} \rangle^2}{\|\mathbf{1}\|^2}} = \sqrt{4n(\alpha - \alpha^2)} = \sqrt{4n(1 - (2\alpha - 1)^2)}$$

Therefore, we get that $\hat{\theta}$ is a $\sqrt{\log n}/\sqrt{4n(1-(2\alpha-1)^2)S}$-consistent estimator. □

References

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

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

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

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

[BFT17] P. Bartlett, D. J. Foster, and M. Telgarsky. "Spectrally-normalized margin bounds for neural networks". In: *arXiv preprint arXiv:1706.08498* (2017).

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

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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

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

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

[Cha05] S. Chatterjee. "Concentration inequalities with exchangeable pairs". PhD thesis. Citeseer, 2005.

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

[Che+20] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li. "Simple and deep graph convolutional networks". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1725–1735.

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

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

[Dag+20] Y. Dagan, C. Daskalakis, N. Dikkala, and A. V. Kandiros. "Learning ising models from one or several samples". In: *arXiv preprint arXiv:2004.09370* (2020).

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

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

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

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

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

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

[Fen+20] W. Feng, J. Zhang, Y. Dong, Y. Han, H. Luan, Q. Xu, Q. Yang, E. Kharlamov, and J. Tang. "Graph Random Neural Networks for Semi-Supervised Learning on Graphs". In: *Advances in Neural Information Processing Systems* 33 (2020).

[Gam03] D. Gamarnik. "Extension of the PAC framework to finite and countable Markov chains". In: *IEEE Transactions on Information Theory* 49.1 (2003), pp. 338–345.

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

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

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

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

[KW16] T. N. Kipf and M. Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).

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

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

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

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

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

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

[Pei+20] H. Pei, B. Wei, K. C.-C. Chang, Y. Lei, and B. Yang. "Geom-gcn: Geometric graph convolutional networks". In: *arXiv preprint arXiv:2002.05287* (2020).

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

[RWY11] G. Raskutti, M. J. Wainwright, and B. Yu. "Minimax rates of estimation for high-dimensional linear regression over l_q -balls". In: *IEEE transactions on information theory* 57.10 (2011), pp. 6976–6994.

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

[SK13] C. Shalizi and A. Kontorovich. "Predictive PAC Learning and Process Decompositions". In: *Advances in Neural Information Processing Systems*. 2013.

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

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

[Ver10] R. Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *arXiv preprint arXiv:1011.3027* (2010).

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

[YCS16] Z. Yang, W. Cohen, and R. Salakhudinov. "Revisiting semi-supervised learning with graph embeddings". In: *International conference on machine learning*. PMLR. 2016, pp. 40–48.

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