Statistical Estimation from Dependent Data

Yuval Dagan [1] **Constantinos Daskalakis** [1] **Nishanth Dikkala** [2] **Surbhi Goel** [3] **Vardis Kandiros** [1]

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

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

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:

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

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

[1]MIT EECS [2]Google Research [3]Microsoft Research NYC. Correspondence to: Yuval Dagan <dagan@mit.edu>, Constantinos Daskalakis <costis@mit.edu>, Nishanth Dikkala <nishanthd@google.com>, Surbhi Goel <surbgoel@microsoft.com>, Vardis Kandiros <kandiros@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

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:

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

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 $y_1, \ldots, y_n$ 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. (Geman & Graffigne, 1986; Ellison, 1993; Felsenstein, 2004; Chatterjee, 2005; Daskalakis et al., 2011; 2017).

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. (2019), 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 & Mukherjee (2018); Bhattacharya & Mukherjee (2018); Chatterjee (2007), 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 $(x_i)_{i=1}^n$, 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 $y_1, \ldots, y_n$ are sampled from (3) conditioning on the feature vectors $x_1, \ldots, x_n$. As explained earlier, in comparison to the standard supervised learning setting without dependencies, the statistical challenge that arises here is that, while the labels $y_1, \ldots, y_n$ 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 $y_1, \ldots, y_n$ are sampled from (3) conditioning on $x_1, \ldots, x_n$, 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.* $\geq 1-\delta$,

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

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. (2019), 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*,*

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

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*.*

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. (2019). 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 (Bartlett et al., 2017). In particular, we assume that the neural network takes the form:

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

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 (Bartlett et al., 2017) 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* $\geq 1-\delta$*,*

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

Notice that, in this case, we do not provide guarantees for the estimation of θ. Since these networks are often over-parametrized, 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. $\geq 1-\delta$,*

$$\frac{1}{n}\sum_{i=1}^{n}(f_{\hat{\theta}}(x_i) - f_{\theta^*}(x_i))^2 \lesssim \mathcal{C}_1(\mathcal{F}, X, \beta^*, \theta^*) \inf_{\epsilon \geq 0}\left(\log\frac{n}{\delta} + \epsilon n + \log N(\mathcal{F}, X, \epsilon)\right), \tag{5}$$

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 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 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 (Glaeser et al., 1996), welfare participation (Bertrand et al., 2000), school achievement (Sacerdote, 2001), retirement plan participation (Duflo & Saez, 2003), and obesity (Christakis & Fowler, 2013; Trogdon et al., 2008). These applications have motivated substantial work in Econometrics (see e.g. Manski (1993); Bramoullé et al. (2009) and their references), where identification results have been pursued and debated, mostly in linear auto-regressive models; see also Daskalakis et al. (2019). In Statistical Learning Theory, learnability and uniform convergence bounds have been shown in the presence of sample dependencies; see e.g. Yu (1994); Gamarnik (2003); Berti et al. (2009); Mohri & Rostamizadeh (2009); Pestov (2010); Mohri & Rostamizadeh (2010); Shalizi & Kontorovich (2013); London et al. (2013); Kuznetsov & Mohri (2015); London et al. (2016); McDonald & Shalizi (2017); Dagan et al. (2019). 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. (2019) 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 (2007); Bhattacharya & Mukherjee (2018); Ghosal & Mukherjee (2018); Dagan et al. (2020), 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 (Daskalakis et al., 2019).

At a higher level, single-sample statistical estimation is both a classical and an emerging field (Besag, 1974; Bresler & Nagaraj, 2018; Chen et al., 2019; Dagan et al., 2020) 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.

[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 $f \in \mathcal{F}$.

2. The Estimation Algorithm

In all our theorems, the estimator we use is the Maximum Pseudo-Likelihood Estimator (MPLE), first proposed by Besag (1974) and defined as follows

$$(\hat{\theta},\hat{\beta}) := \arg\max_{\theta,\beta} \prod_{i=1}^{n} \mathbb{P}_{\theta,\beta}[y_i|x,y_{-i}] \equiv \prod_{i=1}^{n} \frac{\exp\left(y_i\left(f_\theta(x_i)+\beta\sum_{j=1}^{n}A_{ij}y_j\right)\right)}{2\cosh\left(f_\theta(x_i)+\beta\sum_{j=1}^{n}A_{ij}y_j\right)}\ , \quad (6)$$

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 (Sly & Sun, 2014). 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 (6) 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 $\|A\|_\infty$, which is defined as $\|A\|_\infty := \max_{1\le i\le n}\sum_{j=1}^{n}|A_{ij}|$. 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 $\ge 1-\delta$:

$$|\hat{\theta}-\theta^*|\lesssim\frac{\sqrt{\log\frac{n}{\delta}}}{\|A\|_F}. \quad (7)$$

In fact, we will show the tighter bound of:

$$|\hat{\theta}-\theta^*|\lesssim\sup_{\lambda\in\mathbb{R}}\frac{\sqrt{\log\frac{n}{\delta}}}{\|\lambda A\|_F+\left\|x-\lambda A\mathbf{tanh}\left(\frac{\beta^* x}{\lambda}+\theta^* x\right)\right\|_2} \quad (8)$$

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 (8), we note that it is stronger than the right hand side of (7). 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 (8). Let $\varphi(\theta,\beta)$ be the negative pseudo log-likelihood for the pair (θ,β), namely, minus the log of the quantity in (6). 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

$$\frac{1}{2}v^\top\nabla^2\varphi(\theta',\beta')v\le-v^\top\nabla\varphi(\theta^*,\beta^*)\le|v^\top\nabla\varphi(\theta^*,\beta^*)|. \quad (9)$$

Using concentration inequalities from (Dagan et al., 2020), we can show that w.pr. $\ge 1-\delta$ (w.r.t. the randomness of the $y_1,\ldots,y_n$ which are implicit arguments of φ), any $u\in\mathbb{R}^2$

satisfies

$$\frac{|u^\top \nabla \varphi(\theta^*, \beta^*)|}{u^\top \nabla^2 \varphi(\theta', \beta') u} \lesssim \frac{\sqrt{\log n/\delta}}{\|u_\beta A\|_F + \|u_\theta x + u_\beta A y\|}. \quad (10)$$

After substituting $u = v$, it follows from (9) that the left hand side of (10) 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

$$|\hat{\theta} - \theta^*| = |v_\theta| \le \frac{\sqrt{\log n/\delta}}{\|\lambda A\|_F + \|x - \lambda A y\|} \le \sup_{\lambda \in \mathbb{R}} \frac{\sqrt{\log n/\delta}}{\|\lambda A\|_F + \|x - \lambda A y\|}. \quad (11)$$

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 (8). 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 (11) by that in (8), which corresponds to showing

$$\|\lambda A\|_F + \|x - \lambda A y\| \gtrsim \|x - \lambda A \mathbf{tanh}((\beta^*/\lambda)x + \theta^* x)\|. \quad (12)$$

We start by describing a rough and informal intuition for proving (12), 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

$$y \approx \mathbf{tanh}(\beta^* A y + \theta^* x). \quad (13)$$

We assume towards contradiction that (12) 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 (13), we obtain that $y \approx \mathbf{tanh}(\beta^* x/\lambda + \theta^* x)$. Making this substitution in $\|x - \lambda A y\|$, we obtain (12).

Now, we will argue more formally about the previous claims to derive (12). Using the triangle inequality, we get

$$\begin{aligned}\|x - \lambda A \mathbf{tanh}(\beta^* x/\lambda + \theta^* x)\| \le \\ \|x - \lambda A y\| + \|\lambda A y - \lambda A \mathbf{tanh}(\beta^* A y + \theta^* x)\| \\ + \|\lambda A \mathbf{tanh}(\beta^* A y + \theta^* x) - \lambda A \mathbf{tanh}(\beta^* x/\lambda + \theta^* x)\|. \end{aligned} \quad (14)$$

We would like to bound each of the three terms on the right hand side by a constant times the left hand side of (12). For the first term, this is trivial. Further, we can show that the third term on the right hand side of (14) is bounded by the first term, using the Lipschitzness of $\tanh$:

$$\begin{aligned}\|\lambda A \mathbf{tanh}(\beta^* A y + \theta^* x) - \lambda A \mathbf{tanh}((\beta^*/\lambda)x + \theta^* x)\| \\ \le \|\lambda A \beta^* A y - A \beta^* x\| \le \|\beta^* A\|_2 \|x - \lambda A y\| \\ \le O(\|x - \lambda A y\|),\end{aligned}$$

where $\|\beta^* A\|_2 \le O(1)$ 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 (13). In order to bound this error term, we use the method of exchangeable pairs developed in (Chatterjee, 2005), which provides a strong and general concentration inequality for non-independent random variables. We can show that with high probability, this term will be $O(\|\lambda \beta^* A\|_2) \le O(\lambda \|A\|_2) \le O(\lambda \|A\|_F)$, since $B = O(1)$. Combining the above bounds we derive (12), 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 $d(\omega, \omega') \le \epsilon$. 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 $h, h' \in \mathbb{R}^n$:

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

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

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

where

$$U_f := \|f_\theta(X) - f_{\theta^*}(X)\|_2^2/n.$$

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 (8) 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 $\mathcal{C}_1 \le \mathcal{C}_1'$ into (5), substituting $f_\theta(x_i) = \theta x_i$ and substituting $\lambda = -(\beta - \beta^*)/(\theta - \theta^*)$, (8) 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. $\ge 1 - \delta$,

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

It is possible to prove that $\mathcal{C}_1' \le O(1/\|A\|_F^2)$, 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

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

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 $\mathcal{H} = \{h^* + tv : t \in \mathcal{T} \subseteq \mathbb{R}\}$. 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. $\ge 1 - \delta$, for all $i \in [N]$,

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

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).

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 $\beta = 0$). This is tantamount to not having an underlying influence network between the nodes. We observe that MPLE-β consistently outperforms MPLE-0 by a significant margin.

Datasets. We utilize three public citation datasets - Cora, Citeseer and Pubmed (Yang et al., 2016). 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. The adjacency information A_{ij} between two nodes is given as a real-value between 0 and 1 in all 3 of the these datasets. Table 1 gives the statistics of the datasets.

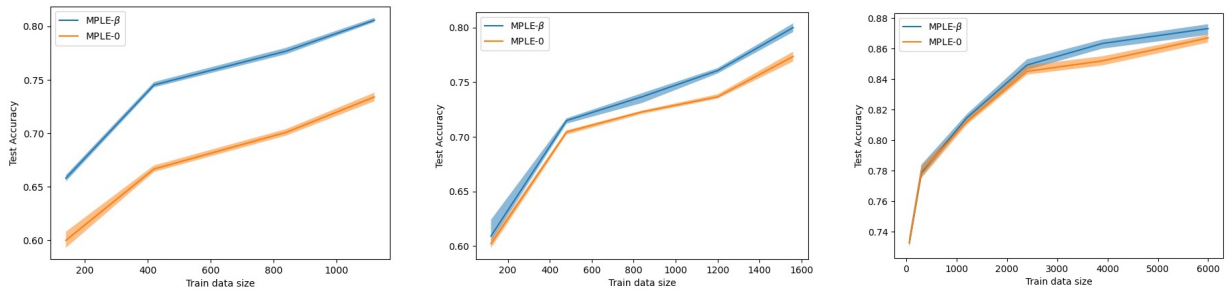

Figure 1. From Left to Right: Plots of the accuracy of MPLE-β (*blue*) vs MPLE-0 (*orange*) for Cora, Citeseer, Pubmed respectively as we increase the training data size gradually while maintaining the class probabilities.

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

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 for a lot of these datasets is graph neural network (GNN) (Chen et al., 2020) 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 (Kipf & Welling, 2016; Feng et al., 2020) 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 et al., 2020). In this setup, we consider 10 random splits of the entire dataset. Each split maintains class distribution by splitting the set of nodes of 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:

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

Note that this extension is a strict generalization of the model we used in our theory (which only deals with binary classification). Even in this more general setting, we observe significant empirical benefits which attests to the applicability of our approach in more general settings than those considered in our theory. Using this we compute the MPLE-β objective.

For both MPLE-0 and MPLE-β, our underlying model $f_\theta : \mathbb{R}^{\#\text{features}} \to \mathbb{R}^{\#\text{classes}}$ 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 (Chen et al., 2020) 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 (Feng et al., 2020).

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

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

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.

References

Bartlett, P., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

Berti, P., Crimaldi, I., Pratelli, L., and Rigo, P. Rate of convergence of predictive distributions for dependent data. *Bernoulli*, 15(4):1351–1367, 2009.

Bertrand, M., Luttmer, E. F., and Mullainathan, S. Network effects and welfare cultures. *The Quarterly Journal of Economics*, 115(3):1019–1055, 2000.

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

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

Bramoullé, Y., Djebbari, H., and Fortin, B. Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55, 2009.

Bresler, G. and Nagaraj, D. Optimal single sample tests for structured versus unstructured network data. *arXiv preprint arXiv:1802.06186*, 2018.

Chatterjee, S. *Concentration inequalities with exchangeable pairs*. PhD thesis, Citeseer, 2005.

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

Chen, J. Y., Valiant, G., and Valiant, P. How bad is worst-case data if you know where it comes from? *arXiv*, abs/1911.03605, 2019.

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

Christakis, N. A. and Fowler, J. H. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577, 2013.

Dagan, Y., Daskalakis, C., Dikkala, N., and Jayanti, S. Learning from weakly dependent data under dobrushin's condition. In *Conference on Learning Theory*, pp. 914–928, 2019.

Dagan, Y., Daskalakis, C., Dikkala, N., and Kandiros, A. V. Learning ising models from one or several samples. *arXiv preprint arXiv:2004.09370*, 2020.

Daskalakis, C., Mossel, E., and Roch, S. Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel's conjecture. *Probability Theory and Related Fields*, 149(1):149–189, 2011.

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

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

Duflo, E. and Saez, E. The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *The Quarterly journal of economics*, 118(3):815–842, 2003.

Ellison, G. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates Sunderland, 2004.

Feng, W., Zhang, J., Dong, Y., Han, Y., Luan, H., Xu, Q., Yang, Q., Kharlamov, E., and Tang, J. Graph random neural networks for semi-supervised learning on graphs. *Advances in Neural Information Processing Systems*, 33, 2020.

Gamarnik, D. Extension of the pac framework to finite and countable markov chains. *IEEE Transactions on Information Theory*, 49(1):338–345, 2003.

Geman, S. and Graffigne, C. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, pp. 1496–1517. American Mathematical Society, 1986.

Ghosal, P. and Mukherjee, S. Joint estimation of parameters in ising model. *Annals of Statistics*, 2018.

Glaeser, E. L., Sacerdote, B., and Scheinkman, J. A. Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548, 1996.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *ICLR*, 2016.

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

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

London, B., Huang, B., and Getoor, L. Stability and generalization in structured prediction. *The Journal of Machine Learning Research*, 17(1):7808–7859, 2016.

Manski, C. F. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

McDonald, D. J. and Shalizi, C. R. Rademacher complexity of stationary sequences. *arXiv preprint arXiv:1106.0730*, 2017.

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

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

Pei, H., Wei, B., Chang, K. C.-C., Lei, Y., and Yang, B. Geom-gcn: Geometric graph convolutional networks. *ICLR*, 2020.

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

Sacerdote, B. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly journal of economics*, 116(2):681–704, 2001.

Shalizi, C. and Kontorovich, A. Predictive pac learning and process decompositions. In *Advances in Neural Information Processing Systems*, 2013.

Sly, A. and Sun, N. Counting in two-spin models on d-regular graphs. *The Annals of Probability*, 42(6):2383–2416, 2014.

Trogdon, J. G., Nonnemaker, J., and Pais, J. Peer effects in adolescent overweight. *Journal of health economics*, 27 (5):1388–1399, 2008.

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

Yu, B. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pp. 94–116, 1994.