Logistic regression with peer-group effects via inference in higher-order Ising models

Constantinos Daskalakis
EECS & CSAIL, MIT
costis@csail.mit.edu

Nishanth Dikkala
EECS & CSAIL, MIT
nishanthd@csail.mit.edu

Ioannis Panageas
ISTD & SUTD
ioannis@sutd.edu.sg

Abstract

Spin glass models, such as the Sherrington-Kirkpatrick, Hopfield and Ising models, are all well-studied members of the exponential family of discrete distributions, and have been influential in a number of application domains where they are used to model correlation phenomena on networks. Conventionally these models have quadratic sufficient statistics and consequently capture correlations arising from pairwise interactions. In this work we study extensions of these to models with higher-order sufficient statistics, modeling behavior on a social network with peer-group effects. In particular, we model binary outcomes on a network as a higher-order spin glass, where the behavior of an individual depends on a linear function of their own vector of covariates and some polynomial function of the behavior of others, capturing peer-group effects. Using a *single*, high-dimensional sample from such model our goal is to recover the coefficients of the linear function as well as the strength of the peer-group effects. The heart of our result is a novel approach for showing strong concavity of the log pseudo-likelihood of the model, implying statistical error rate of $\sqrt{d/n}$ for the Maximum Pseudo-Likelihood Estimator (MPLE), where d is the dimensionality of the covariate vectors and n is the size of the network (number of nodes). Our model generalizes vanilla logistic regression as well as the models studied in recent works of [14, 24, 18], and our results extend these results to accommodate higher-order interactions.

1 Introduction

Did you choose red rather than blue because some inherent attributes of yours biased you towards red, or because your social environment biased you towards that color? Of course, the answer is typically "both." Indeed, a long literature in econometrics and the social sciences has substantiated the importance of peer effects in network behavior in topics as diverse as criminal activity (see e.g. [25]), welfare participation (see e.g. [1]), school achievement (see e.g. [32]), participation in retirement plans (see e.g. [20]), and obesity (see e.g. [34, 16]). On the other hand, estimating the mechanisms through which peer and individual effects drive behavior in such settings has been quite challenging; see e.g. [29, 3].

From a modeling perspective, a class of probabilistic models that are commonly used to model binary behavior in social networks are spin glass models, such as the well-studied Sherrington-Kirkpatrick, Hopfield and Ising models. In these models, a vector of binary behaviors $\mathbf{y} \in \{-1, 1\}^V$ across all nodes of some network $G = (V, E)$ is sampled jointly according to the Gibbs distribution, $p(\mathbf{y}) = \frac{1}{Z} \exp(-\mathrm{En}(\mathbf{y}))$, defined by some energy function $\mathrm{En}(\mathbf{y})$ of the aggregate behavior, where the functional form of $\mathrm{En}(\cdot)$ typically depends on characteristics of the nodes as well as the structure of their social network. Such models studied originally in Statistical Physics, have found myriad applications in diverse fields, including Probability Theory, Markov Chain Monte Carlo, Computer Vision, Computational Biology, Game Theory, and, related to our focus, Economics and the Social Sciences [28, 11, 22, 19, 23, 21, 30].

Closely related to our work, a series of recent works have studied estimation of spin glass models incorporating both peer and individual effects as drivers of behavior [12, 24, 18]. Generalizing the classical logistic regression model, these works consider models of binary behavior on a network, conforming to the following general class of models. Suppose that the nodes of a social network $G = (V, E)$ have individual

characteristics $\mathbf{x}_i \in \mathbb{R}^d$, $i \in V$, and sample binary behaviors $\mathbf{y} \in \{\pm 1\}^V$ according to some measure that combines individual and peer effects, taking the following form:

$$\Pr[\mathbf{y}] = \frac{1}{Z_{\theta,\beta}} \exp\left(\sum_{i \in V} (\theta^\top \mathbf{x}_i) y_i + \beta \cdot f(\mathbf{y})\right), \tag{1}$$

where a linear function $\theta^\top \mathbf{x}_i$ of node i's individual characteristics determines the "external field" on that node, i.e. the direction and strength of the "local push" of that node towards -1 or $+1$, and some function $f(\mathbf{y})$ of the nodes' joint behavior expresses what configurations in $\{\pm 1\}^V$ are encouraged by peer-group effects. In particular, setting $\beta = 0$ recovers the standard logistic regression model, where nodes choose their behaviors independently, but setting $\beta > 0$ incorporates peer-group effects, as expressed by f. Without loss of generality, f is a multi-linear function, and we can take E to contain a hyperedge for each monomial in f, i.e. take $f(\mathbf{y}) = \sum_{\mathbf{e} \in E} w_{\mathbf{e}} \mathbf{y}_{\mathbf{e}}$ where $\mathbf{y}_{\mathbf{e}} = \prod_{i \in \mathbf{e}} y_i$.

Given a collection $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ of covariates, some function $f : \{\pm 1\}^V \to \mathbb{R}$, and a *single* sample $\mathbf{y}$ drawn from a model conforming to (1), the afore-cited works of Chatterjee [12], Ghosal and Mukerjee [24] and Daskalakis et al. [18] provide computationally and statistically efficient algorithms for estimating θ and β. Specifically, these works study the restriction of model (1) to the case where f contains only pair-wise effects, i.e. where function f is a multilinear function of degree 2. In particular, Chatterjee [12] studies the case where $\theta = 0$ and f is bilinear, Ghosal and Mukerjee [24] the case where $d = 1$, all x_i's equal 1, and f is bilinear, while Daskalakis et al. [18] the general bilinear case. Extending these works, the goal of our work is to provide computationally and statistically efficient estimation methods for models where f has peer effects of higher-order. As such, our new methods can accommodate richer models, capturing a much broader range of social interactions, e.g. settings where nodes belong in various groups, and dislike fragile majorities in the groups they belong to. Our main result is the following.

Theorem 1.1 (Informal)**.** *Let $G = (V, E, w : E \to \mathbb{R})$ be a weighted hypergraph with edges of cardinality at least two and at most some constant m, and let $f(\mathbf{y}) = \sum_{\mathbf{e} \in E} w_{\mathbf{e}} \mathbf{y}_{\mathbf{e}}$. Assume that each vertex has bounded degree (Assumption 2.1) and the hypergraph is dense enough (Assumption 2.2). Moreover, assume that the true parameters θ_0, β_0 and the feature vectors have bounded ℓ_2 norm, and the empirical covariance matrix of the feature vectors has singular values upper and lower bounded by constants (Assumption 2.3). Then, there exists a polynomial-time algorithm, which, given a single sample from model* (1)*, outputs an estimate $(\tilde{\theta}, \tilde{\beta})$ such that $\left\|(\tilde{\theta}, \tilde{\beta}) - (\theta_0, \beta_0)\right\|_2$ is $O\left(\sqrt{\frac{d}{n}}\right)$, with probability at least* 99%*, where $n = |V|$.*

Discussion of Main Result. First, let us discuss the assumptions made in our statement. Note that the assumptions about θ and the x_i's are standard, and are commonly made even for vanilla logistic regression without peer effects ($\beta = 0$). The assumption about the boundedness of β and the degree of the hypergraph is needed so that the peer-group effects do not overwhelm the individual effects, making θ non-identifiable. Finally, the assumption on the density of the hypergraph is needed so that the individual effects do not overwhelm the peer-group effects, making β non-identifiable. Our assumptions about β and the hypergraph are generalizations of corresponding assumptions made in prior work. As such, our main result is a direct generalization of prior work to accommodate higher-order peer effects.

We should also discuss the importance, in both our work and the work we build upon [12, 24, 18], of estimating the parameters of our model using a *single* sample, which stands in contrast to other recent work studying estimation of Ising models and more general Markov Random Fields from multiple samples; see e.g. [6, 4, 7, 35, 27, 8]. The importance of estimating from a single sample arises from the applications motivating our work, where it is more common than not that we really only have a single sample of node behavior across the whole network, and cannot obtain a fresh independent sample of behavior tomorrow or within a reasonable time-frame.

Techniques. Towards obtaining Theorem 1.1, we encounter several technical challenges. A natural approach is to use our single sample to perform Maximum Likelihood Estimation. However, this approach faces

two important challenges. First, it has been shown that the single sample Maximum Likelihood Estimator is not necessarily consistent [12]. Second, the likelihood function involves the partition function $Z_{\theta,\beta}$, which is generally computationally intractable to compute. In view of these issues, we follow instead the approach followed in prior work. Rather than maximizing the likelihood of the sample, we maximize its *pseudolikelihood*, defined as $\prod_i \Pr[y_i \mid \mathbf{y}_{-i}]$. This concave function of our parameters θ and β is computationally easy to optimize, however we need to show that its maximum is consistent. To argue this we establish two main properties of the log-pseudolikelihood: (i) the log-pseudolikelihood is strongly concave in the neighborhood of its maximum; and (ii) its gradient at the true model parameters is bounded. As both the Hessian and the gradient of log-pseudolikelihood are functions of the vector of variables $\mathbf{y}$, which are jointly sampled, to argue (i) and (ii) we need to control functions of dependent random variables. To do this we use exchangeable pairs, adapting the technique of [13], combined with a parity argument on G and f's partial derivatives. In turn, (i) and (ii) suffice to establish the consistency of the Maximum Pseudolikelihood Estimator (MPLE).

1.1 More Related Work

Learning and testing questions on Ising models have been widely studied in diverse contexts. A popular instantiation of the learning problem is structure learning, where given access to multiple i.i.d. samples from the model we wish to infer the underlying graph's structure. This was first studied for tree graphical models by [15] and has since then seen a lot of work both in terms of upper bounds and lower bounds side [33]. More recently, [5] gave a striking algorithm for structure learning in bounded degree graphs which required samples only logarithmic in the number of nodes of the graph. The running time and sample complexity of this approach was improved in later works of [35, 27, 26]. The works of [27, 26] provide learning results for MRFs with higher-order interactions on alphabet of sizes larger than 2. Property testing questions on Ising models have also been studied by [17]. All of the above works, however, make use of access to many independent samples from a Ising model. Closer to the model we consider in this paper is the line of work initiated by [14] and extensions in the works of [2, 24, 18] wherein we try to infer an Ising model described by a few parameters using a single sample from the model. [9, 31] study hypothesis testing questions on the Ising model from a single sample.

2 Preliminaries

We use bold letters such as $\mathbf{x}, \mathbf{y}$ to denote vectors and capital letters A, W to denote matrices. All vectors are assumed to be column vectors, i.e. $\dim \times 1$ (except when we refer to the parameters as (θ, β) instead of $(\theta^\top, \beta)$). We will refer to W_{ij} as the $(i,j)^{th}$ entry of matrix W. We will use the following matrix norms. For a $n \times n$ matrix W,

$$\|W\|_2 = \max_{\|x\|_2=1} \|Wx\|_2 , \;\; \|W\|_\infty = \max_{j\in[n]} \sum_{i=1}^{n} |W_{ij}| , \;\; \|W\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}^2}. \tag{2}$$

When W is a symmetric matrix we have that $\|W\|_2 \le \|W\|_\infty \le \|W\|_F \le \sqrt{n}\,\|W\|_2 \le \sqrt{n}\,\|W\|_\infty$ and in general we have $\|W\|_2^2 \le \|W\|_\infty \|W\|_1$.

We use λ to denote eigenvalues of a matrix and σ to denote its singular values. $\lambda_{\min}$ refers to the smallest eigenvalue and $\lambda_{\max}$ to the largest, and similar notation is used for the singular values. We use $\mathbf{e}$ or a collection $\{z_1, ..., z_m\}$ to denote a hyperdge and moreover its weight is denoted by $w_{\mathbf{e}}$ or $w_{(z_1,...,z_m)}$.

We will say an estimator $\hat{\theta}_n$ is consistent with a rate $r(n)$ (or equivalently $r(n)$-consistent) with respect to the true parameter θ_0 if there exists an integer n_0 and a constant $C > 0$ such that for every $n > n_0$, with probability at least 99%,

$$\left\|\hat{\theta}_n - \theta_0\right\|_2 \le \frac{C}{r(n)}.$$

2.1 Ising Model and Inference

The Ising model is a well-studied binary graphical model. We provide the description of the model here.

1. **Ising Model (simple):** Given a weighted undirected graph $G(V, E)$ with $|V| = n$ and a $n \times n$ weight matrix W and assignment $\sigma : V \to \{-1, +1\}$, an Ising model is the following probability distribution on the 2^n configurations of σ:

$$\Pr\{\mathbf{y} = \sigma\} = \frac{\exp\left(\sum_{v \in V} h_v \sigma_v + \beta \sigma^\top W \sigma\right)}{Z_G} \tag{3}$$

   where

$$Z_G = \sum_{\tilde{\sigma}} \exp\left(\sum_{v \in V} h_v \tilde{\sigma}_v + \beta \tilde{\sigma}^\top W \tilde{\sigma}\right)$$

   is the partition function of the system (or renormalization factor). Moreover the term $\sum_v h_v \sigma_v$ is called the external field and β is called the inverse temperature. It can be observed that, without loss of generality, we can restrict the matrix W to have zeros on its diagonal.

2. **Ising Model (Hypergraph):** Given a hypergraph graph $G(V, E)$ (each edge $\mathbf{e}$ has at most m incident vertices and at least two), weights $w_{\mathbf{e}}$ and assignment $\sigma : V \to \{-1, +1\}$, an Ising model is the following probability distribution on the 2^n configurations of σ:

$$\Pr\{\mathbf{y} = \sigma\} = \frac{\exp\left(\sum_{v \in V} h_v \sigma_v + \beta f(\sigma)\right)}{Z_G}, \tag{4}$$

   where $f(\sigma) = \sum_{\mathbf{e} \in E(G)} w_{\mathbf{e}} \sigma_{\mathbf{e}}$ and $\sigma_{\mathbf{e}} = \prod_{v \in \mathbf{e}} \sigma_v$. Observe that $f(\sigma)$ is a multilinear polynomial of degree m (since $y_v^2 = 1$ for all v and every realization, weighted hypergraphs capture all distributions with f a polynomial function).

Inference of Ising models with Hypergraphs: In this paper we focus on the following modification of the Ising model for hypergraphs. It is assumed that we are given **one sample** from the following distribution:

$$\Pr[\mathbf{y} = \sigma] = \frac{\exp(\beta f(\sigma) + \sum_v (\mathbf{x}_v^\top \theta) \sigma_v)}{Z_G(\beta, \theta)},$$

where β, θ are unknown parameters, $f : \{-1, +1\}^n \to \mathbb{R}$ is a polynomial (multilinear) function and each summand is of degree at most m and at least two ($Z_G(\beta, \theta)$ is the renormalization factor again). The goal is to *estimate* the parameters β and θ. This problem is a generalization of the logistic regression model with dependent observations problem as appeared in [18] (for $m = 2$), applied to hypergraphs.

- Observe that for each index v we can write $f(\mathbf{y}) = y_v f_v(\mathbf{y}_{-v}) + f_{-v}(\mathbf{y}_{-v})$ (both f_{-v}, f_v are multilinear functions that do not depend on y_v). It is easy to see that $f_v(\mathbf{y}_{-v}) = \frac{\partial f}{\partial y_v}$. Each hyperedge $\mathbf{e}$ is a collection of at most m vertices $v \in V$. One may write $\mathbf{y}_{\mathbf{e}} = \prod_{v \in \mathbf{e}} y_v$ and moreover $f(\mathbf{y}) = \sum_{\mathbf{e} \in E} w_{\mathbf{e}} \mathbf{y}_{\mathbf{e}}$ and $y_v f_v(\mathbf{y}_{-v}) = \sum_{\mathbf{e} \in E, v \in \mathbf{e}} w_{\mathbf{e}} y_{\mathbf{e}}$.

- For all vertices v and $\sigma_v \in \{\pm 1\}$, conditioning on a realization of the response variables $\mathbf{y}_{-v}$:

$$\Pr[y_v = \sigma_v] = \frac{1}{1 + \exp\left(-2\left(\theta^\top \mathbf{x}_v + \beta f_v(\mathbf{y}_{-v})\right) \sigma_v\right)}. \tag{5}$$

- Interpretation: The probability that the conditional distribution of y_v assigns to $+1$ is determined by the logistic function applied to $2\left(\theta^\top \mathbf{x}_v + \beta f_v(\mathbf{y}_{-v})\right)$ instead of $2\theta^\top \mathbf{x}_v$.

2.2 Assumptions

Our Assumptions can be listed below:

Assumption 2.1 (Bounded degree)**.**

$$\sum_{\mathbf{e}:i\in\mathbf{e}} |w_{\mathbf{e}}| \leq 1, \tag{6}$$

for all vertices i, where $\mathbf{e}$ captures the hyperedges. The number one on the R.H.S can be replaced with any constant. This assumption is mainly used in our concentration bounds.

Assumption 2.2 (Enough weight at the hyperedges)**.**

$$\sum_{\substack{\mathbf{e}\in E,\\ |e|=m}} w_e^2 \text{ is } \Omega(n), \tag{7}$$

This assumption is mainly used to prove strong concavity of the pseudolikelihood for the estimation of β.

Assumption 2.3 (Parameters and features)**.** *The true parameter β_0 belongs in some interval $(-B, B)$ and $\|\theta_0\|_2 < \Theta$ for some known constants B, Θ that are independent of n, d. We denote by $\mathbb{B} \subseteq \mathbb{R}^{d+1}$, $\mathbb{B} = \{(\theta, \beta) \in \mathbb{R}^{d+1}, |\beta| \leq B, \|\theta\|_2 \leq \Theta\}$ (i.e., the closure of the set that the parameters may belong to).*

Moreover for every feature vector $\mathbf{x}_v$ we have $\|\mathbf{x}_v\|_2 \leq M$ (for some known constant M independent of n, d). Finally, the covariance matrix (of size $d \times d$) of the feature vectors, i.e., $\frac{1}{n}X^\top X$ where $X^\top = (\mathbf{x}_1\ \mathbf{x}_2 \dots \mathbf{x}_n)$ has minimum and maximum eigenvalues bounded by constants (independent of n, d) and the projection matrix $F = I - X(X^\top X)^{-1}X^\top$ satisfies $\|F\|_\infty$ is bounded by a constant (one without loss of generality).

2.3 Pseudo-Likelihood - Gradient and Hessian

The pseudolikelihood as defined by Chatterjee in [14] for a simpler model and instantiated in our model is given by the following expression:

$$PL(\theta, \beta) := \left(\prod_{i=1}^{n} \Pr[y_i | \mathbf{y}_{-i}]\right)^{1/n} = \left(\prod_{i=1}^{n} \frac{\exp((\theta^\top \mathbf{x}_i + \beta f_i(\mathbf{y}_{-i}))y_i)}{\exp(\theta^\top \mathbf{x}_i + \beta f_i(\mathbf{y}_{-i})) + \exp(-\theta^\top \mathbf{x}_i - \beta f_i(\mathbf{y}_{-i}))}\right)^{1/n} \tag{8}$$

Taking the log, the log pseudolikelihood for a specific sample $\mathbf{y}$ is given by:

$$LPL(\theta, \beta) := \frac{1}{n}\sum_{i=1}^{n} \left[y_i \beta f_i(\mathbf{y}_{-i}) + y_i(\theta^\top \mathbf{x}_i) - \ln\cosh(\beta f_i(\mathbf{y}_{-i}) + \theta^\top \mathbf{x}_i)\right] - \ln 2, \tag{9}$$

The first order conditions give:

$$\begin{array}{l} \frac{\partial LPL(\theta,\beta)}{\partial \beta} = \frac{1}{n}\sum_{i=1}^{n} \left[y_i f_i(\mathbf{y}_{-i}) - f_i(\mathbf{y}_{-i}) \tanh(\beta f_i(\mathbf{y}_{-i}) + \theta^\top \mathbf{x}_i)\right] = 0, \\ \frac{\partial LPL(\theta,\beta)}{\partial \theta_k} = \frac{1}{n}\sum_{i=1}^{n} \left[y_i x_{i,k} - x_{i,k} \tanh(\beta f_i(\mathbf{y}_{-i}) + \theta^\top \mathbf{x}_i)\right] = 0. \end{array} \tag{10}$$

The solution to equation (10) is called Maximum Pseudolikelihood Estimator (Hessian is negative semidefinite, see below) and is denoted by $(\hat{\theta}, \hat{\beta})$ or $(\hat{\theta}_{MPL}, \hat{\beta}_{MPL})$.

The Hessian $H_{(\theta,\beta)}$ of the log-pseudolikelihood is given by:

$$\begin{array}{l} \frac{\partial^2 LPL(\theta,\beta)}{\partial \beta^2} = -\frac{1}{n}\sum_{i=1}^{n} \frac{f_i^2(\mathbf{y}_{-i})}{\cosh^2(\beta f_i(\mathbf{y}_{-i}) + \theta^\top \mathbf{x}_i)}, \\ \frac{\partial^2 LPL(\theta,\beta)}{\partial \beta \partial \theta_k} = -\frac{1}{n}\sum_{i=1}^{n} \frac{x_{i,k} f_i(\mathbf{y}_{-i})}{\cosh^2(\beta f_i(\mathbf{y}_{-i}) + \theta^\top \mathbf{x}_i)}, \\ \frac{\partial^2 LPL(\theta,\beta)}{\partial \theta_l \partial \theta_k} = -\frac{1}{n}\sum_{i=1}^{n} \frac{x_{i,l} x_{i,k}}{\cosh^2(\beta f_i(\mathbf{y}_{-i}) + \theta^\top \mathbf{x}_i)}. \end{array} \tag{11}$$

Writing the Hessian differently we get

$$H_{(\theta,\beta)} = -\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\cosh^2(\beta f_i(\mathbf{y}_{-i}) + \theta^\top \mathbf{x}_i)} X_i X_i^\top$$

where $X_i = (\mathbf{x}_i^\top, f_i(\mathbf{y}_{-i}))^\top$. Thus $-H$ is a positive semidefinite matrix and LPL is concave. Moreover if (θ, β) satisfies Assumptions 2.1 and 2.3 it follows that

$$\tfrac{1}{\cosh^2(B+M\cdot\Theta)} \cdot \left(\tfrac{1}{n}\sum_{i=1}^{n} X_i X_i^\top\right) \preceq -H_{(\theta,\beta)} \preceq \left(\tfrac{1}{n}\sum_{i=1}^{n} X_i X_i^\top\right). \tag{12}$$

Remark 2.1 (LPL is smooth)**.** *Since* $\|X_i\|_2^2 = \|\mathbf{x}_i\|_2^2 + f_i^2(\mathbf{y}_{-i}) \le \Theta^2 + 1$ *(assuming Assumption 2.1 trivially holds* $|f_i(\mathbf{y}_{-i})| \le 1$*) it holds that* $\lambda_{\max}(-H_{(\theta,\beta)}) \le \Theta^2 + 1$ *for all* $(\theta, \beta) \in \mathbb{R}^{d+1}$ *which satisfy Assumption 2.3, hence* $-LPL$ *is a* $\Theta^2 + 1$*-smooth function, i.e.* $-\nabla LPL$ *is* $\Theta^2 + 1$*-Lipschitz.*

We conclude this session with an important lemma that explains the reason we need the technical lemmas in Section 3 and involves that gradient and the Hessian of the log-psudolikelihood (appeared in [18]).

Lemma 2.1 (Consistency of the MPLE [18])**.** *Let* (θ_0, β_0) *be the true parameter. We define* $(\theta_t, \beta_t) = (1-t)(\theta_0, \beta_0) + t(\hat{\theta}_{MPL}, \hat{\beta}_{MPL})$ *and let* $\mathcal{D} \in [0, 1]$ *be the largest value such that* $(\theta_{\mathcal{D}}, \beta_{\mathcal{D}}) \in \mathbb{B}$ *(if it does not intersect the boundary of* $\mathbb{B}$*, then* $\mathcal{D} = 1$*), where* $\mathbb{B}$ *is defined in Assumption 2.3. Then,*

$$\begin{aligned}\|\nabla LPL(\theta_0, \beta_0)\|_2 &\ge \mathcal{D} \min_{(\theta,\beta)\in\mathbb{B}} \lambda_{\min}\left(-H_{(\theta,\beta)}\right) \left\|(\theta_0 - \hat{\theta}_{MPL}, \beta_0 - \hat{\beta}_{MPL})\right\|_2 \\ &= \min_{(\theta,\beta)\in\mathbb{B}} \lambda_{\min}\left(-H_{(\theta,\beta)}\right) \|(\theta_0 - \theta_{\mathcal{D}}, \beta_0 - \beta_{\mathcal{D}})\|_2 .\end{aligned}$$

To prove the main result, we apply Lemma 2.1 by showing: (in the rest of the paper)

1. *A concentration result* for $\|\nabla LPL(\theta_0, \beta_0)\|_2^2$ around d/n (Section 3.1) which in words gives that the gradient of the log-pseudolikelihood at the true parameter is small (note that it is zero at the MPLE) (I).
2. *A lower bound (positive constant that depends on the degree of polynomial f)* for $\min_{(\theta,\beta)\in\mathbb{B}} \lambda_{\min}\left(-H_{(\theta,\beta)}\right)$ (Section 3.2) with high probability (II).

We combine the above with the observation that $\mathcal{D} = 1$ for n sufficiently large. This is true because $\|(\theta_{\mathcal{D}} - \theta_0, \beta_{\mathcal{D}} - \beta_0)\|_2 \to 0$ as $n \to \infty$ (is of order $\frac{1}{\sqrt{n}}$ and that any point on the boundary of $\mathbb{B}$ has a fixed (independent of n) positive distance to (θ_0, β_0) since (θ_0, β_0) lies in the interior of $\mathbb{B}$.

This gives the desired rate of consistency which we show in Section 3.2.

3 Maximum Pseudo-Likelihood (MPLE): Concentration and Strong Concavity

In this section, we prove Theorem 1.1. In words, we show consistency of the MPLE which we prove via bullets (I), (II) and then applying Lemma 2.1 as stated in the previous section. Our main result is formally given below:

Theorem 3.1 (Main (Formal))**.** *Consider the model of (1) with Assumptions 2.1, 2.2, 2.3 and denote Maximum Pseudo-Likelihood Estimate (MPLE) with* $(\hat{\theta}_{MPL}, \hat{\beta}_{MPL})$*. With probability* 99.9% *it holds that*

$$\left\|(\hat{\theta}_{MPL}, \hat{\beta}_{MPL}) - (\theta_0, \beta_0)\right\|_2 \le O\left(\sqrt{\frac{d}{n}}\right) 2^{O(m)}$$

and we can compute an estimate with the same order of consistency in $O(\ln n)$ *iterations of projected gradient descent (Algorithm in Section B) where each iteration takes polynomial (in* n*) time.*

3.1 Concentration Results for Gradient (I)

The first main technical Lemma is to show that the norm of the gradient of the log-pseudolikelihood is small enough at the true parameters (Corollary 3.1). This is necessary because we are working with the finite sample pseudolikelihood (empirical). In what follows we show that the difference between sum of $y_i f(y_{-i})$ (or $y_i \mathbf{x}_i$) and the sum of their conditional expectations is small.

Lemma 3.1 (Variance Bound 1)**.** *It holds that*

$$\mathbb{E}_{\theta_0,\beta_0}\left[\left(\sum_{i=1}^{n} y_i f_i(\mathbf{y}_{-i}) - f_i(\mathbf{y}_{-i})\tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)\right)^2\right] \leq (12+4B)(m-1)n.$$

Lemma 3.2 (Variance Bound 2)**.** *It holds that*

$$\mathbb{E}_{\theta_0,\beta_0}\left[\sum_{k=1}^{d}\left(\sum_{i=1}^{n} x_{i,k}y_i - x_{i,k}\tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)\right)^2\right] \leq (1+B)4M^2\cdot(m-1)dn.$$

We are now ready to prove bullet (I).

Corollary 3.1. *For each $\delta > 0$ and n sufficiently large, with probability $1-\delta$ it holds that*

$$\Pr_{\theta_0,\beta_0}\left[\|\nabla LPL(\theta_0,\beta_0)\|_2 \leq C\sqrt{\frac{1}{\delta}}\sqrt{\frac{d}{n}}\right],$$

for some global constant C.

Proof. Observe that (see Equations of the gradient, left-hand side in (10))

$$\begin{aligned}\|\nabla LPL(\theta_0,\beta_0)\|_2^2 \quad &= \tfrac{1}{n^2}\textstyle\sum_{k=1}^{d}\left(\sum_{i=1}^{n} x_{i,k}y_i - x_{i,k}\tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)\right)^2 \\ &+\tfrac{1}{n^2}\left(\textstyle\sum_{i=1}^{n} y_i f_i(\mathbf{y}_{-i}) - f_i(\mathbf{y}_{-i})\tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)\right)^2.\end{aligned} \tag{13}$$

The claim is an application of Lemmas 3.1, 3.2 and Markov's inequality. □

3.2 Strong Concavity of log-Pseudolikelihood (II)

Schur's complement. Let

$$X^\top = (\mathbf{x}_1\ \mathbf{x}_2 \dots \mathbf{x}_n),$$

which is the matrix of the covariates (of size $d \times n$). Using Equation (12) (the negative Hessian of log-Pseudolikelihood dominates the matrix below) we get

$$-H \succeq \frac{1}{\cosh^2(B + M\cdot\Theta)} G \text{ where } G := \begin{pmatrix} \frac{1}{n}X^\top X & \frac{1}{n}X^\top \mathbf{f} \\ \frac{1}{n}\mathbf{f}^\top X & \frac{1}{n}\|\mathbf{f}\|_2^2 \end{pmatrix},$$

and $\mathbf{f} := (f_1(\mathbf{y}_{-1}), ..., f_n(\mathbf{y}_{-n}))$.

We set $Q = \frac{1}{n}X^\top X$ and use the properties of Schur complement on the matrix

$$G - \lambda I = \begin{pmatrix} Q - \lambda I & \frac{1}{n}X^\top \mathbf{f} \\ \frac{1}{n}\mathbf{f}^\top X & \frac{1}{n}\|\mathbf{f}\|_2^2 - \lambda \end{pmatrix}$$

to get that

$$\det(G-\lambda I) = \det(Q-\lambda I)\det\left(\frac{1}{n}\mathbf{f}^\top\left(I - \frac{1}{n}X\left(Q-\lambda I\right)^{-1}X^\top\right)\mathbf{f} - \lambda\right). \tag{14}$$

Therefore the minimum eigenvalue of G is at least a positive constant as long as the minimum eigenvalues of

$$Q \text{ and } \frac{1}{n}\mathbf{f}^\top \left(I - \frac{1}{n}XQ^{-1}X^\top\right)\mathbf{f}$$

are at least positive constants independent of n, d. Recall from our assumptions (Assumption 2.3) we have that $\lambda_{\min}(Q) \geq c_1$ always where c_1 is a positive constant independent of n, d. Hence, it remains to show that

$$\lambda_{\min}\left(\frac{1}{n}\mathbf{f}^\top \left(I - \frac{1}{n}XQ^{-1}X^\top\right)\mathbf{f}\right) \geq c_2$$

for a positive constant c_2 with high probability (with respect to the randomness in drawing $\mathbf{y}$).

Denoting $F = I - X(X^\top X)^{-1}X^\top = I - \frac{1}{n}XQ^{-1}X^\top$, observe that F has the property $F^2 = F$ (i.e. is idempotent) and hence all the eigenvalues of F are $0, 1$ (since is of rank $n-d$, it has d eigenvalues zero and $n-d$ eigenvalues one). Our goal is to show that

Lemma 3.3.

$$\mathbf{f}^\top F\mathbf{f} = \|F\mathbf{f}\|_2^2 \geq c_2 n \quad \textit{with probability } 1 - o(1), \tag{15}$$

where the probability is with respect to the randomness in drawing $\mathbf{y}$.

Lower bound on the "expectation". Our first key lemma, is to prove a lower bound on the conditional expectation of each summand of the quantity $\|F\mathbf{f}\|_2^2 = \sum_i (F\mathbf{f})_i^2$ which is captured in Corollary 3.2 and is a consequence of the lemma below.

Lemma 3.4 (Parity Lemma)**.** *Fix a sequence of indices $z_1, ..., z_{m-1}$ and an index i. It holds that*

$$\mathbb{E}_{\theta_0,\beta_0}[(F\mathbf{f})_i^2 | \mathbf{y}_{-z_1,...,-z_{m-1}}] \geq \frac{e^{-(B+M\cdot\Theta)(m-1)}}{2^{m-1}} \left(\sum_j F_{ij} w_{j,z_1,...,z_{m-1}}\right)^2.$$

In case $j = z_t$ for some $t < m$ then $w_{j,z_1,...,z_{m-1}} = 0$.

Proof.

$$\begin{aligned}
&\mathbb{E}_{\theta_0,\beta_0}[(F\mathbf{f})_i^2 | \mathbf{y}_{-z_1,...,-z_{m-1}}] = \mathbb{E}_{\theta_0,\beta_0}\left[\left(\sum_j F_{ij} f_j(\mathbf{y}_{-j})\right)^2 \Big| \mathbf{y}_{-z_1,...,-z_{m-1}}\right] \\
&= \mathbb{E}_{\theta_0,\beta_0}\left[\left(\sum_j F_{ij} \sum_{\mathbf{e}, j\in\mathbf{e}} w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j\}}\right)^2 \Big| \mathbf{y}_{-z_1,...,-z_{m-1}}\right] \\
&= \mathbb{E}_{\theta_0,\beta_0}\left[\left(\sum_j F_{ij} \sum_{\mathbf{e}: j, z_1\in\mathbf{e}} w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{\mathbf{j}\}} + \sum_j F_{ij} \sum_{\mathbf{e}: j\in\mathbf{e}, z_1\notin\mathbf{e}} w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j\}}\right)^2 \Big| \mathbf{y}_{-z_1,...,-z_{m-1}}\right] \\
&= \mathbb{E}_{\theta_0,\beta_0}\left[\left(y_{z_1}\sum_{j\neq z_1}\sum_{\mathbf{e}: j, z_1\in\mathbf{e}} F_{ij} w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j,z_1\}} + \sum_j \sum_{\mathbf{e}: j\in\mathbf{e}, z_1\notin\mathbf{e}} F_{ij} w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j\}} + F_{iz_1} f_{z_1}\right)^2 \Big| \mathbf{y}_{-z_1,...,-z_{m-1}}\right].
\end{aligned}$$

It is clear that the square above is at least $(\sum_j \sum_{\mathbf{e}:j,z_1\in\mathbf{e}} F_{ij} w_{\mathbf{e}} \mathbf{y}_{\mathbf{e}\setminus\{j,z_1\}})^2$ depending on $y_{z_1} = \pm 1$. Thus using the fact that $|f_{z_1}| \leq 1$ we conclude that

$$\mathbb{E}_{\theta_0,\beta_0}[(F\mathbf{f})_i^2 | \mathbf{y}_{-\mathbf{e}}] \geq \frac{e^{-(B+M\cdot\Theta)}}{2} \mathbb{E}_{\theta_0,\beta_0}\left[\left(\sum_j \sum_{\mathbf{e}:j,z_1\in\mathbf{e}} F_{ij} w_{\mathbf{e}} \mathbf{y}_{\mathbf{e}\setminus\{j,z_1\}}\right)^2 \Big| \mathbf{y}_{-z_1,\ldots,-z_{m-1}}\right] \tag{16}$$

$$= \frac{e^{-(B+M\cdot\Theta)}}{2} \mathbb{E}_{\theta_0,\beta_0}\left[\left(\sum_j \sum_{\mathbf{e}:j,z_1\in\mathbf{e}} F_{ij} w_{\mathbf{e}} \mathbf{y}_{\mathbf{e}\setminus\{j,z_1\}}\right)^2 \Big| \mathbf{y}_{-z_1,\ldots,-z_{m-1}}\right]. \tag{17}$$

Now observe that (and since $\frac{\partial f_{z_1}}{\partial y_{z_1}} = 0$) $\sum_{j\neq z_1} \sum_{\mathbf{e}:j,z_1\in\mathbf{e}} F_{ij} w_{\mathbf{e}} \mathbf{y}_{\mathbf{e}\setminus\{j,z_1\}} = \sum_j F_{ij} \frac{\partial f_j}{\partial y_{z_1}}$ hence we conclude that

$$\mathbb{E}_{\theta_0,\beta_0}[(F\mathbf{f})_i^2 \big| \mathbf{y}_{-z_1,\ldots,-z_{m-1}}] \geq \frac{e^{-(B+M\cdot\Theta)}}{2} \mathbb{E}_{\theta_0,\beta_0}[(F\tilde{\mathbf{f}})_i^2 \big| \mathbf{y}_{-z_1,\ldots,-z_{m-1}}], \tag{18}$$

where $\tilde{\mathbf{f}} = (\frac{\partial f_1}{\partial y_{z_1}}, ..., \frac{\partial f_n}{\partial y_{z_1}})$. By an induction argument we may conclude that

$$\mathbb{E}_{\theta_0,\beta_0}[(F\mathbf{f})_i^2 \big| \mathbf{y}_{-z_1,\ldots,-z_{m-1}}] \geq \left(\frac{e^{-(B+M\cdot\Theta)}}{2}\right)^{m-1} \mathbb{E}_{\theta_0,\beta_0}[(F\hat{\mathbf{f}})_i^2 \big| \mathbf{y}_{-z_1,\ldots,-z_{m-1}}], \tag{19}$$

where $\hat{f} = (\frac{\partial^{m-1} f_1}{\partial y_{z_1}\ldots\partial y_{z_{m-1}}}, ..., \frac{\partial^{m-1} f_n}{\partial y_{z_1}\ldots\partial y_{z_{m-1}}}) = (\frac{\partial^m f}{\partial y_1 \partial y_{z_1}\ldots\partial y_{z_{m-1}}}, ..., \frac{\partial^m f}{\partial y_n \partial y_{z_1}\ldots\partial y_{z_{m-1}}}) = (w_{1,z_1,\ldots,z_{m-1}}, ..., w_{n,z_1,\ldots,z_{m-1}})$. □

Corollary 3.2 (Tower property)**.** *For each vertex i and distinct vertices $v, z_1, ..., z_{m-2}$ (note i is not necessarily different from v) it holds*

$$\mathbb{E}_{\theta_0,\beta_0}[(F\mathbf{f})_i^2 | \mathbf{y}_{-v}] \geq \frac{e^{-(B+M\cdot\Theta)(m-1)}}{2^{m-1}} \left(\sum_j F_{ij} w_{\{v,z_1,\ldots,z_{m-2}\}\cup\{j\}}\right)^2.$$

Proof. It follows by applying Lemma 3.4 and the tower property. □

In what follows, we define an "adjacency" matrix A which enables us to reduce the general degree m polynomial case to the case where $m = 2$.

Reduction to "simple graphs". To prove strong concavity of the Hessian of the log-pseudolikelihood, we need to show that $\|F\mathbf{f}\|_2^2$ is at least $c_1 n$ with high probability. To do this, we reduce the general problem to the case $m = 2$ by defining the appropriate matrix below and then use the machinery of [18] to show that $\|F\mathbf{f}\|_2^2$ is concentrated around its conditional expectation (see Lemma 3.8).

Let A be the following $n \times n$ matrix: For each column i of A, let

$$(z_1^*, ..., z_{m-2}^*) = \text{argmax}_{z_1,\ldots,z_{m-2}} \left\|(w_{(z_1,\ldots,z_{m-2},i,1)}, ..., w_{(z_1,\ldots,z_{m-2},i,n)})\right\|_2.$$

The j-th entry of column i of A is given by $w_{(z_1^*,\ldots,z_{m-2}^*,i,j)}$. Intuitively, matrix A induces a subgraph of the original hypergraph G. Nevertheless, matrix A contains "enough edges" to infer θ_0, β_0.

Lemma 3.5 (A has big Frobenius norm)**.** *There exists a constant C such that*

$$\|A\|_F^2 \geq Cn. \tag{20}$$

Proof. Define the matrix B of size $|E(G)_m| \times n$ where $E(G)_m$ is the set of edges of cardinality m, $B_{\mathbf{e},i} = w_e \times \mathbf{1}_{i \in \mathbf{e}}$ and $\mathbf{e} \in E(G)_m$. It holds that $\|B\|_F^2$ is $\Omega(n)$. Consider the maximum entry in absolute value per column of B, let b_i, i.e., $b_i = \left\|B^i\right\|_\infty$. Since $\left\|B^i\right\|_1 \leq 1$ (bounded degree assumption) by Holder's inequality we get that $b_i \geq \left\|B^i\right\|_2^2$. Therefore we conclude that $\sum_i b_i \geq \|B\|_F^2$, thus it is $\Omega(n)$. From Cauchy-Schwarz we get that $\sum_i b_i^2 \geq \frac{(\sum_i b_i)^2}{n} \geq \frac{\|B\|_F^4}{n}$ which is $\Omega(n)$.

The proof is complete by observing that $\left\|A^i\right\|_2^2 \geq b_i^2$ for all i and thus $\|A\|_F^2$ is $\Omega(n)$. □

Moreover, A satisfies the bounded degree condition and this is captured by the lemma below.

Lemma 3.6 (Bounding $\|A\|_\infty, \|A\|_1$)**.** *It holds that*

$$\|A\|_1, \|A\|_\infty \leq m - 1.$$

Proof. Each entry in A_{ij} is some weight of an edge that contains i, j (if there exists one otherwise zero). Hence in every row/column, each edge appears at most $m-1$ times and by the bounded degree assumption the claim follows. □

Note that from Lemma 3.5 and 3.6 we get $\|FA\|_F^2$ is also $\Omega(n)$. This is true, since $\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty} \leq m-1$ (Lemma 3.6), thus $\|FA\|_F^2 \geq \|A\|_F^2 - d(m-1)^2$. To proceed, we use a selection index procedure that appeared in [18] (we mention it below for completeness) and which will be useful in the later part of the proof.

An Index Selection Procedure [18]: Given a matrix W, we define $h : [n] \to [n]$ as follows. Consider the following iterative process. At time $t = 0$, we start with the $n \times n$ matrix, $W^1 = W$. At time step t we choose from W^t the row with maximum ℓ_2 norm (let i_t the index of that row, ties broken arbitrarily) and also let $j_t = \text{argmax}_j |W^t_{i_t j}|$ (again ties broken arbitrarily). We set $h(i_t) = j_t$ and W^{t+1} is W^t by setting zeros the entries of i_t^{th} row and column j_t^{th}. We run the process above for n steps to define the **bijection** h. The following lemma is taken from [18].

Lemma 3.7 ([18])**.** *Assume that $\|FA\|_\infty \leq c'_\infty$ and[1] $\|FA\|_F^2 \geq c_F n$ for some positive constant c_∞, c_F and $\|A\|_2, \|A\|_\infty, \|A\|_1$ are also bounded. We run the process described above on FA and get the function h. There exists a constant C (depends on c_F, c_∞) such that*

$$\sum_i |(FA)_{ih(i)}|^2 \geq Cn.$$

Combining Corollary 3.2 (summing over all i) with Lemma 3.7, there exists a constant C (independent of n, d) such that the following inequality is true (always)

$$\sum_i \mathbb{E}_{\theta_0,\beta_0}[(F\mathbf{f})_i^2 | \mathbf{y}_{-h(i)}] \geq \frac{e^{-(B+M\cdot\Theta)(m-1)}}{2^{m-1}} \times \sum_i (FA)_{ih(i)}^2 \geq C \times \frac{e^{-(B+M\cdot\Theta)(m-1)}}{2^{m-1}} \times n. \tag{21}$$

Equation (21) gives us the linear in n lower bound that we want for the sum of conditional expectations of the terms $(F\mathbf{f})_i^2$. Finally we need to show that the term $\sum_i (F\mathbf{f})_i^2$ is not far from $\sum_i \mathbb{E}_{\theta_0,\beta_0}[(F\mathbf{f})_i^2 | \mathbf{y}_{-h(i)}]$ with high probability, thus it is also at least linear in n and Lemma 3.3 would follow. This is captured in the following lemma.

Lemma 3.8 (Bounding the "conditional" variance)**.** *It holds that*

$$\mathbb{E}_{\theta_0,\beta_0}\left[\left(\sum_{i=1}^n (F\mathbf{f})_i^2 - \sum_{i=1}^n \mathbb{E}_{\theta_0,\beta_0}\left[(F\mathbf{f})_i^2 | \mathbf{y}_{-h(i)}\right]\right)^2\right] \leq (80n + 16Bn)(m-1).$$

[1] Recall $F = I - X(X^\top X)^{-1} X^\top$.

Putting it all together

Proof of Theorem 3.1. We can prove now our main result, the approach is similar to [18]. From Corollary 3.1 we get that (for some constant C_1)

$$\Pr\left[\|\nabla LPL(\theta_0,\beta_0)\|_2^2 \leq \frac{C_1 d}{n\delta}\right] \geq 1-\delta. \tag{22}$$

for any constant δ. Next, we have from Lemma 3.3 and the analysis in the beginning of Section 3.2 that, $\min_{(\theta,\beta)\in\mathbb{B}} \lambda_{\min}\left(-H_{(\theta,\beta)}\right) \geq C_2$ for some constant C_2 independent of n, d. Plugging into Lemma 2.1, we get that

$$\|(\theta_{\mathcal{D}}-\theta_0,\beta_{\mathcal{D}}-\beta_0)\|_2 = \mathcal{D}\left\|(\hat{\theta}_{MPL}-\theta_0,\hat{\beta}_{MPL}-\beta_0)\right\|_2 \leq \frac{\|\nabla LPL(\theta_0,\beta_0)\|_2}{\min_{(\theta,\beta)\in\mathbb{B}} \lambda_{\min}\left(-H_{(\theta,\beta)}\right)} \tag{23}$$

Now we have from the above that $\|(\theta_{\mathcal{D}}-\theta_0,\beta_{\mathcal{D}}-\beta_0)\|_2 \to 0$ as $n\to\infty$ and also holds that $\|(\theta_{\mathcal{D}}-\theta_0,\beta_{\mathcal{D}}-\beta_0)\|_2 \to 0$ which implies that $\mathcal{D}=1$ for sufficiently large n. Therefore

$$(23) \implies \left\|(\hat{\theta}_{MPL}-\theta_0,\hat{\beta}_{MPL}-\beta_0)\right\|_2 \leq \frac{\|\nabla LPL(\theta_0,\beta_0)\|_2}{\min_{(\theta,\beta)\in\mathbb{B}} \lambda_{\min}\left(-H_{(\theta,\beta)}\right)} \tag{24}$$

$$\leq O\left(\sqrt{\frac{d}{n}}\right) \tag{25}$$

with probability $\geq 1-\delta$. The analysis of Projected Gradient Descent can be found in the appendix. □

4 Conclusion

In this paper, we focused on the problem of parameter estimation from one sample of a high dimensional discrete distribution that can be viewed as an instantiation Logistic Regression from dependent observations or Inference on Ising models, with high-order peer effects. There are many open questions, we state a few:

- In the consistency rate, there is an exponential dependence on the degree m of the polynomial function f (m now is considered a constant number). Can this be improved?
- Analyze more complicated settings where function f is Lipschitz.

5 Acknowledgements

Constantinos Daskalakis and Nishanth Dikkala were supported by NSF Awards IIS-1741137, CCF-1617730 and CCF-1901292, by a Simons Investigator Award, by the DOE PhILMs project (No. DE-AC05-76RL01830), by the DARPA award HR00111990021, by a Google Faculty award, by the MIT Frank Quick Faculty Research and Innovation Fellowship, and an MIT-IBM Watson AI Lab research grant. Ioannis Panageas was supported by SRG ISTD 2018 136, NRF-NRFFAI1-2019-0003 and NRF2019NRF-ANR2019.

References

[1] Marianne Bertrand, Erzo FP Luttmer, and Sendhil Mullainathan. Network effects and welfare cultures. *The Quarterly Journal of Economics*, 115(3):1019–1055, 2000.

[2] Bhaswar B Bhattacharya, Sumit Mukherjee, et al. Inference in ising models. *Bernoulli*, 24(1):493–525, 2018.

[3] Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55, 2009.

[4] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pages 771–782, New York, NY, USA, 2015. ACM.

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

[6] Guy Bresler, David Gamarnik, and Devavrat Shah. Structure learning of antiferromagnetic Ising models. In *Advances in Neural Information Processing Systems 27*, NIPS '14, pages 2852–2860. Curran Associates, Inc., 2014.

[7] Guy Bresler and Mina Karzand. Learning a tree-structured Ising model in order to make predictions. *arXiv preprint arXiv:1604.06749*, 2016.

[8] Guy Bresler, Frederic Koehler, and Ankur Moitra. Learning restricted boltzmann machines via influence maximization. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019, Phoenix, AZ, USA, June 23-26, 2019.*, pages 828–839, 2019.

[9] Guy Bresler and Dheeraj Nagaraj. Optimal single sample tests for structured versus unstructured network data. *arXiv preprint arXiv:1802.06186*, 2018.

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

[11] Sourav Chatterjee. *Concentration Inequalities with Exchangeable Pairs*. PhD thesis, Stanford University, June 2005.

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

[13] Sourav Chatterjee and Amir Dembo. Nonlinear large deviations. *Advances in Mathematics*, 299:396–450, 2016.

[14] Sourav Chatterjee et al. Estimation in spin glasses: A first step. *The Annals of Statistics*, 35(5):1931–1946, 2007.

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

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

[17] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, Philadelphia, PA, USA, 2018. SIAM.

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

[19] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. 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.

[20] Esther Duflo and Emmanuel Saez. 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.

[21] Glenn Ellison. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

[22] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates Sunderland, 2004.

[23] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, pages 1496–1517. American Mathematical Society, 1986.

[24] Promit Ghosal and Sumit Mukherjee. Joint estimation of parameters in ising model. *arXiv preprint arXiv:1801.06570*, 2018.

[25] Edward L Glaeser, Bruce Sacerdote, and Jose A Scheinkman. Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548, 1996.

[26] Linus Hamilton, Frederic Koehler, and Ankur Moitra. Information theoretic properties of markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems*, pages 2463–2472, 2017.

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

[28] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

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

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

[31] Rajarshi Mukherjee, Sumit Mukherjee, Ming Yuan, et al. Global testing against sparse alternatives under ising models. *The Annals of Statistics*, 46(5):2062–2093, 2018.

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

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

[34] Justin G Trogdon, James Nonnemaker, and Joanne Pais. Peer effects in adolescent overweight. *Journal of health economics*, 27(5):1388–1399, 2008.

[35] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

A Missing Proofs

Proof of Lemma 3.1. We use the powerful technique of exchangeable pairs as introduced by Chatterjee and employed by Chatterjee and Dembo. First it holds by assumption that it trivially follows that $|f_i(\mathbf{y}_{-i})| \leq 1$ for all i and $\mathbf{y}_{-i} \in \{-1,+1\}^{n-1}$. Set

$$Q(\mathbf{y}) := \sum_i (y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)) f_i(\mathbf{y}_{-i}), \tag{26}$$

hence we get

$$\frac{\partial Q(\mathbf{y})}{\partial y_j} = \sum_i \left(\mathbf{1}_{i=j} - \frac{\beta_0 \frac{\partial f_i(\mathbf{y}_{-i})}{\partial y_j}}{\cosh^2(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)} \right) f_i(\mathbf{y}_{-i}) + \left(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)\right) \frac{\partial f_i(\mathbf{y}_{-i})}{\partial y_j}. \tag{27}$$

We will bound the absolute value of each summand. First observe that $\left|\frac{\partial f_i(\mathbf{y}_{-i})}{\partial y_j}\right| \leq \sum_{\mathbf{e}: i,j \in \mathbf{e}} |w_\mathbf{e}|$, hence we can bound the second term as follows

$$\left| (y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)) \frac{\partial f_i(\mathbf{y}_{-i})}{\partial y_j} \right| \leq 2 \sum_{\mathbf{e}: i,j \in e} |w_\mathbf{e}|. \tag{28}$$

Using the fact that $\frac{1}{\cosh^2(x)} \leq 1$ it also follows that

$$\begin{aligned} \left| \sum_i \left(\mathbf{1}_{i=j} - \frac{\beta_0 \frac{\partial f_i(\mathbf{y}_{-i})}{\partial y_j}}{\cosh^2(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)} \right) f_i(\mathbf{y}_{-i}) \right| &\leq |f_j(\mathbf{y}_{-j})| + \textstyle\sum_{i \neq j} \sum_{\mathbf{e}: i,j \in \mathbf{e}} |\beta_0| |w_\mathbf{e} f_i(\mathbf{y}_{-i})| \\ &\leq |f_j(\mathbf{y}_{-j})| + \textstyle\sum_{i \neq j} \sum_{\mathbf{e}: i,j \in \mathbf{e}} |\beta_0| |w_\mathbf{e}|. \end{aligned} \tag{29}$$

Using (28) and (29) it follows that $\left|\frac{\partial Q(\mathbf{y})}{\partial y_j}\right| \leq \sum_{i \neq j} \sum_{\mathbf{e}: i,j \in e} |w_\mathbf{e}|(2 + |\beta_0|) + |f_j(\mathbf{y}_{-j})|$. Finally let $\mathbf{y^j} = (\mathbf{y}_{-j}, -1)$ and note that

$$|Q(\mathbf{y}) - Q(\mathbf{y^j})| \leq 2 \cdot \left(\sum_{i \neq j} \sum_{\mathbf{e}: i,j \in e} |w_\mathbf{e}|(2 + |\beta_0|) + \max_{\mathbf{y}_{-j}} |f_j(\mathbf{y}_{-j})| \right) \tag{30}$$

$$\leq 2 \cdot (1 + (2 + B)(m-1) \sum_{\mathbf{e}: j \in \mathbf{e}} |w_\mathbf{e}|) \tag{31}$$

$$\leq (4 + 2B)(m-1) + 2 \leq (6 + 2B)(m-1). \tag{32}$$

We have all the ingredients to complete the proof. We first observe that

$$\sum_i \mathbb{E}_{\theta_0, \beta_0}[(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)) Q(\mathbf{y^i}) f_i(\mathbf{y}_{-i})] = 0, \tag{33}$$

since

$$\begin{aligned} &\mathbb{E}_{\theta_0, \beta_0}[(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)) Q(\mathbf{y^i}) f_i(\mathbf{y}_{-i})] = \\ &= \mathbb{E}_{\theta_0, \beta_0}[\mathbb{E}[(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)) Q(\mathbf{y^i}) f_i(\mathbf{y}_{-i}) | \mathbf{y}_{-i}]] = 0. \end{aligned} \tag{34}$$

Therefore it follows

$$\begin{aligned}\mathbb{E}_{\theta_0,\beta_0}[Q^2(\mathbf{y})] &= \mathbb{E}_{\theta_0,\beta_0}\left[Q(\mathbf{y})\cdot\left(\sum_i (y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))f_i(\mathbf{y}_{-i})\right)\right] \\ &= \mathbb{E}_{\theta_0,\beta_0}\left[\sum_i \left(Q(\mathbf{y})(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))f_i(\mathbf{y}_{-i})\right)\right] \\ &= \sum_i \mathbb{E}_{\theta_0,\beta_0}\left[(Q(\mathbf{y}) - Q(\mathbf{y}^{\mathbf{i}}))\cdot(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))f_i(\mathbf{y}_{-i})\right] \\ &\leq \sum_i 2\cdot(6+2B)(m-1) = (12+4B)(m-1)n.\end{aligned}$$

□

Proof of Lemma 3.2. We fix a coordinate k and set

$$Q(\mathbf{y}) := \sum_i (y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))x_{i,k}, \tag{35}$$

hence we get $\frac{\partial Q(\mathbf{y})}{\partial y_j} = \sum_i \left(\mathbf{1}_{i=j} - \frac{\beta_0 \frac{\partial f_i(\mathbf{y}_{-i})}{\partial y_j}}{\cosh^2(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i)}\right) x_{i,k}$. We will bound the term as follows

$$\left|\frac{\partial Q(\mathbf{y})}{\partial y_j}\right| \leq |x_{j,k}| + \sum_{i\neq j} |\beta_0||x_{i,k}| \sum_{\mathbf{e}: i,j\in\mathbf{e}} |w_{\mathbf{e}}|. \tag{36}$$

Finally let $\mathbf{y}^{\mathbf{j}} = (\mathbf{y}_{-j}, -1)$ and note that

$$|Q(\mathbf{y}) - Q(\mathbf{y}^{\mathbf{j}})| \leq 2\cdot\left(|x_{j,k}| + \sum_{i\neq j} |\beta_0||x_{i,k}| \sum_{\mathbf{e}: i,j\in\mathbf{e}} |w_{\mathbf{e}}|\right). \tag{37}$$

We have all the ingredients to complete the proof. We first observe that

$$\sum_i \mathbb{E}_{\theta_0,\beta_0}[(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))Q(\mathbf{y}^{\mathbf{i}})x_{i,k}] = 0, \tag{38}$$

since

$$\begin{aligned}&\mathbb{E}_{\theta_0,\beta_0}[(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))Q(\mathbf{y}^{\mathbf{i}})x_{i,k}] = \\ &= \mathbb{E}_{\theta_0,\beta_0}[\mathbb{E}[(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))Q(\mathbf{y}^{\mathbf{i}})x_{i,k}|\mathbf{y}_{-i}]] = 0.\end{aligned} \tag{39}$$

Therefore it follows

$$\begin{aligned}
\mathbb{E}_{\theta_0,\beta_0}[Q^2(\mathbf{y})] &= \mathbb{E}_{\theta_0,\beta_0}\left[Q(\mathbf{y})\cdot\left(\sum_i (y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))x_{i,k}\right)\right] \\
&= \mathbb{E}_{\theta_0,\beta_0}\left[\sum_i \left(Q(\mathbf{y})(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))x_{i,k}\right)\right] \\
&= \sum_i \mathbb{E}_{\theta_0,\beta_0}\left[(Q(\mathbf{y}) - Q(\mathbf{y}^{\mathbf{i}}))\cdot(y_i - \tanh(\beta_0 f_i(\mathbf{y}_{-i}) + \theta_0^\top \mathbf{x}_i))x_{i,k}\right] \\
&\leq \sum_i 4\cdot(x_{i,k}^2 + |x_{i,k}|\sum_{j\neq i}|\beta_0||x_{j,k}|\sum_{\mathbf{e}:i,j\in\mathbf{e}}|w_{\mathbf{e}}|) \\
&\leq 4\sum_i |x_{i,k}|^2 + B|x_{i,k}|\max_j |x_{j,k}|\sum_{j\neq i}\sum_{\mathbf{e}:i,j\in\mathbf{e}}|w_{\mathbf{e}}| \\
&\leq 4M^2 n + \sum_i 4BM^2(m-1)\sum_{\mathbf{e}:i\in\mathbf{e}}|w_{\mathbf{e}}| = 4nM^2(1+B(m-1)) \\
&\leq 4n(m-1)M^2(1+B),
\end{aligned}$$

and the claim follows by summing over all the coordinates. □

Proof of Lemma 3.8. For each i, we expand the term $\mathbb{E}_{\theta_0,\beta_0}\left[(F\mathbf{f})_i^2|\mathbf{y}_{-h(i)}\right]$ and we get $\mathbb{E}_{\theta_0,\beta_0}\left[(F\mathbf{f})_i^2|\mathbf{y}_{-h(i)}\right] = \mathbb{E}_{\theta_0,\beta_0}\left[\left(y_{h(i)}\sum_{j\neq h(i)}\sum_{\mathbf{e}:j,h(i)\in\mathbf{e}}F_{ij}w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j,h(i)\}} + \sum_j\sum_{\mathbf{e}:j\in\mathbf{e},h(i)\notin\mathbf{e}}F_{ij}w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j\}} + F_{ih(i)}f_{h(i)}\right)^2|\mathbf{y}_{-h(i)}\right]$. We set $z_{it}(\mathbf{y}) = 2\left(\sum_{j\neq t}\sum_{\mathbf{e}:j,t\in\mathbf{e}}F_{ij}w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j,t\}}\right)\left(\sum_j\sum_{\mathbf{e}:j\in\mathbf{e},t\notin\mathbf{e}}F_{ij}w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j\}} + F_{it}f_t\right)$ (does not depend on y_t) and we get that the expectation we need to bound is equal to

$$\mathbb{E}_{\theta_0,\beta_0}\left[\left(\sum_i z_{ih(i)}(\mathbf{y})y_{h(i)} - z_{ih(i)}(\mathbf{y})\tanh\left(\beta_0 f_{h(i)}(\mathbf{y}) + \theta_0^\top \mathbf{x}_{h(i)}\right)\right)^2\right].$$

First it holds that $\frac{\partial z_{it}}{y_j} = 2\left(\sum_{j'\neq t}\sum_{\mathbf{e}:j',j,t\in\mathbf{e}}F_{ij'}w_{\mathbf{e}}y_{\mathbf{e}\setminus\{t,j,j'\}}\right)\left(\sum_{j'}\sum_{\mathbf{e}:j'\in\mathbf{e},t\notin\mathbf{e}}F_{ij'}w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j'\}} + F_{it}f_t\right) +$ $+2\left(\sum_{j'\neq t}\sum_{\mathbf{e}:j',t\in\mathbf{e}}F_{ij'}w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j',t\}}\right)\left(\sum_{j'}\sum_{\mathbf{e}:j',j\in\mathbf{e},t\notin\mathbf{e}}F_{ij'}w_{\mathbf{e}}\mathbf{y}_{\mathbf{e}\setminus\{j',j\}} + F_{it}\frac{\partial f_t}{y_t}\right)$ and $\frac{\partial z_{it}}{y_t} = 0$. Also by the bounded degree condition it holds that $|z_{it}| \leq 4$ as long as $\|F\|_\infty$ is bounded by one. The rest of the proof follows as in Lemma 3.7 in [18] (using exchangeable pairs). □

B Projected Gradient Descent

The following is a well-known fact for Projected Gradient Descent (Theorem 3.10 from [10]).

Theorem B.1. *Let f be α-strongly convex and λ-smooth on compact set $\mathcal{X}$. Then projected gradient descent with stepsize $\eta = \frac{1}{\lambda}$ satisfies for $t \geq 0$*

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \leq e^{-\frac{\alpha t}{\lambda}}\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2. \tag{40}$$

Therefore, setting $R = \|\mathbf{x}_1 - \mathbf{x}^*\|_2$ and by choosing $t = \frac{2\lambda\ln\frac{R}{\epsilon}}{\alpha}$ it is guaranteed that $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2 \leq \epsilon$.

We consider the function $LPL(\theta,\beta)$ (log-pseudolikelihood as defined in Section 2.3) and we would like to approximate $(\hat{\theta},\hat{\beta})$ within $\frac{1}{\sqrt{n}}$ in ℓ_2 distance. The stepsize in Theorem B.1 should be $\eta = \frac{1}{\Theta^2+1}$ by Remark

2.1.

ALGORITHM 1: Projected Gradient Descent

Data: Vector sample $\mathbf{y}$, "Magnetizations" $f_i(\mathbf{y}_{-i}) = y_i \sum_{\mathbf{e}: i \in \mathbf{e}} w_{\mathbf{e}} y_{\mathbf{e}}$, Feature vectors $\mathbf{x}_i$
Result: Maximum Pseudolikelihood Estimate

$\beta^0 = 0, \theta^0 = \mathbf{0}, \text{normgrad} = +\infty$, $\eta = \frac{1}{\Theta^2+1}$;
$t = 0$;
while *normgrad* $> \frac{1}{\sqrt{n}}$ **do**
  $\text{grad}_\theta = 0$;
  $\text{grad}_\beta = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i f_i(\mathbf{y}) - f_i(\mathbf{y}) \tanh(\beta^t f_i(\mathbf{y}) + \theta^{t\top} \mathbf{x}_i) \right]$;
  for $k = 1; k \leq d; k++$ **do**
    $\text{grad}_{\theta_k} = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i x_{i,k} - x_{i,k} \tanh(\beta^t f_i(\mathbf{y}) + \theta^{t\top} \mathbf{x}_i) \right]$;
    $\text{grad}_\theta = \text{grad}_\theta + \text{grad}_{\theta_k}^2$;
  end
  $\text{normgrad} = \sqrt{\text{grad}_\beta^2 + \text{grad}_\theta}$;
  $\beta^{t+1} = \beta^t - \eta \text{grad}_\beta$ % update β^t;
  for $k = 1; k \leq d; k++$ **do**
    $\theta_k^{t+1} = \theta_k^t - \eta \text{grad}_{\theta_k}$ % update θ_k^t;
  end
  % ℓ_2 projection
  if $\beta^{t+1} < -B$ **then**
    $\beta^{t+1} = -B$;
  end
  if $\beta^{t+1} > B$ **then**
    $\beta^{t+1} = B$;
  end
  $\text{norm}_\theta = 0$;
  for $k = 1; k \leq d; k++$ **do**
    $\text{norm}_\theta = \text{norm}_\theta + (\theta_k^{t+1})^2$;
  end
  if $\sqrt{norm_\theta} > \Theta$ **then**
    $\theta^{t+1} = \theta^{t+1} \frac{\Theta}{\sqrt{\text{norm}_\theta}}$;
  end
  $t = t + 1$;
end
return (θ^t, β^t)