Journal of Machine Learning Research 25 (2024) 1-62 Submitted 9/22; Revised 4/24; Published 5/24

High Dimensional Logistic Regression Under Network Dependence

Somabha Mukherjee SOMABHA@NUS.EDU.SG
Department of Statistics and Data Science
National University of Singapore, Singapore

Ziang Niu ZIANGNIU@WHARTON.UPENN.EDU
Department of Statistics and Data Science
University of Pennsylvania, Philadelphia, USA

Sagnik Halder SHALDER@UFL.EDU
Department of Statistics
University of Florida, Gainesville, USA

Bhaswar B. Bhattacharya BHASWAR@WHARTON.UPENN.EDU
Department of Statistics and Data Science
University of Pennsylvania, Philadelphia, USA

George Michailidis GMICHAIL@UFL.EDU
Department of Statistics and Data Science
University of California, Los Angeles, Los Angeles, USA

Editor: Ji Zhu

Abstract

Logistic regression is a key method for modeling the probability of a binary outcome based on a collection of covariates. However, the classical formulation of logistic regression relies on the independent sampling assumption, which is often violated when the outcomes interact through an underlying network structure, such as over a temporal/spatial domain or on a social network. This necessitates the development of models that can simultaneously handle both the network 'peer-effect' (arising from neighborhood interactions) and the effect of (possibly) high-dimensional covariates. In this paper, we develop a framework for incorporating such dependencies in a high-dimensional logistic regression model by introducing a quadratic interaction term, as in the Ising model, designed to capture the pairwise interactions from the underlying network. The resulting model can also be viewed as an Ising model, where the node-dependent external fields linearly encode the high-dimensional covariates. We propose a penalized maximum pseudo-likelihood method for estimating the network peer-effect and the effect of the covariates (the regression coefficients), which, in addition to handling the high-dimensionality of the parameters, conveniently avoids the computational intractability of the maximum likelihood approach. Under various standard regularity conditions, we show that the corresponding estimate attains the classical high-dimensional rate of consistency. In particular, our results imply that even under network dependence it is possible to consistently estimate the model parameters at the same rate as in classical (independent) logistic regression, when the true parameter is sparse and the underlying network is not too dense. Consequently, we derive the rates of consistency of our proposed estimator for various natural graph ensembles, such as bounded degree graphs, sparse Erdős-Rényi random graphs, and stochastic block models. We also develop

an efficient algorithm for computing the estimates and validate our theoretical results in numerical experiments. An application to selecting genes in clustering spatial transcriptomics data is also discussed.

Keywords: High-dimensional inference, Ising models, logistic regression, Markov random fields, network data, penalized regression, pseudo-likelihood, random graphs.

1. Introduction

Logistic regression (Hosmer et al., 2013; McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972) is a very popular and widely used method for modeling the probability of a binary response based on multiple features/predictor variables. It is a mainstay of modern multivariate statistics that has found widespread applications in various fields, including machine learning, biological and medical sciences, economics, marketing and finance industries, and social sciences. For example, in machine learning it is regularly used for image classification and in the medical sciences it is often used to predict the risk of developing a particular disease based on the patients' observed characteristics, among others. To describe the model formally, denote the vector of predictor variables (covariates) by $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N \in \mathbb{R}^d$ and the independent response variables by $X_1, \ldots, X_N \in \{-1, 1\}$. Then, the logistic regression model for the probability of a positive outcome conditional on the covariates is given by

$$\mathbb{P}(X_i = 1 | \boldsymbol{Z}_i) = \frac{e^{\boldsymbol{\theta}^\top \boldsymbol{Z}_i}}{e^{\boldsymbol{\theta}^\top \boldsymbol{Z}_i} + e^{-\boldsymbol{\theta}^\top \boldsymbol{Z}_i}},$$

for $1 \leqslant i \leqslant N$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)^\top \in \mathbb{R}^d$ is the vector of regression coefficients.[1] Using the independence of the response variables, the joint distribution of $\boldsymbol{X} := (X_1, \ldots, X_N)$ given $\boldsymbol{Z} := (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N) \in \mathbb{R}^{d \times N}$ can be written as:

$$\mathbb{P}(\boldsymbol{X} | \boldsymbol{Z}) = \prod_{i=1}^{N} \frac{e^{X_i \boldsymbol{\theta}^\top \boldsymbol{Z}_i}}{e^{\boldsymbol{\theta}^\top \boldsymbol{Z}_i} + e^{-\boldsymbol{\theta}^\top \boldsymbol{Z}_i}} = \frac{1}{\mathcal{Z}_N(\boldsymbol{\theta}, \boldsymbol{Z})} \exp\left\{ \sum_{i=1}^{N} X_i (\boldsymbol{\theta}^\top \boldsymbol{Z}_i) \right\}, \tag{1}$$

where $\mathcal{Z}_N(\boldsymbol{\theta}, \boldsymbol{Z}) = \prod_{i=1}^{N} \frac{1}{e^{\boldsymbol{\theta}^\top \boldsymbol{Z}_i} + e^{-\boldsymbol{\theta}^\top \boldsymbol{Z}_i}}$ is the normalizing constant. It is well-known from the classical theory of generalized linear models that the parameter $\boldsymbol{\theta}$ in (1) can be estimated at rate $O(1/\sqrt{N})$ for fixed dimensions, using the maximum likelihood (ML) method (see, for example, Lehmann and Romano (2005); McCullagh and Nelder (1989); Vaart (1998)).

The classical framework of logistic regression described above is, however, inadequate if the independence assumption on the response variables is violated, which is often the case if the observations are collected over a temporal or spatial domain or on a social network. The recent accumulation of dependent network data in modern applications has accentuated the need to develop realistic and mathematically tractable methods for modeling high-dimensional distributions with underlying network dependencies (*network peer-effect*). Towards this, the Ising model, initially proposed in statistical physics to model ferromagnetism (Ising, 1925), has turned out to be a useful primitive for modeling such datasets, which arise naturally in spatial statistics, social networks, epidemic modeling, computer vision, neural networks, and computational biology, among others (see Banerjee et al. (2015);

1. Note that we are parameterizing the outcomes as $\{-1, 1\}$ instead of the more standard $\{0, 1\}$ to integrate this within the framework of the Ising model (defined in (2)), where the $\{-1, 1\}$ notation is more common.

Geman and Graffigne (1986); Green and Richardson (2002); Hopfield (1982); Montanari and Saberi (2010) and references therein). This can be viewed as a discrete exponential family with binary outcomes, wherein the sufficient statistic is a quadratic form designed to capture correlations arising from pairwise interactions. Formally, given an interaction matrix $A := ((a_{ij}))_{1\leqslant i,j\leqslant N}$ and binary vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_N) \in \mathcal{C}_N = \{-1, 1\}^N$, the Ising model with parameters β and h encodes the dependence among the coordinates of $\boldsymbol{X}$ as follows:

$$\mathbb{P}_{\beta,h}(\boldsymbol{X}) = \frac{1}{2^N \mathcal{Z}_N(\beta, h)} \exp\left\{\beta \sum_{1\leqslant i<j\leqslant N} a_{ij} X_i X_j + h \sum_{i=1}^{N} X_i\right\}, \tag{2}$$

where the *normalizing constant* $\mathcal{Z}_N(\beta, h)$ is determined by the condition $\sum_{\boldsymbol{X}\in\mathcal{C}_N} \mathbb{P}_{\beta,h}(\boldsymbol{X}) = 1$ (so that $\mathbb{P}_{\beta,h}$ is a probability measure). In statistical physics the parameters β and h are referred to as the *inverse temperature* and the *magnetic field*, respectively.

This paper is motivated by applications where in addition to peer effects, arising from an underlying network structure, there are other variables (covariates) associated with the nodes of the network, which affect the outcome of the response variables. For example, in the data collected by the National Longitudinal Study (Harris, 2007) students in grades 7–12 were asked to name up to 10 friends and answer many questions about age, gender, race, socio-economic background, personal and school life, and health, where it becomes imperative to model the peer-effect and the effect of the covariates simultaneously. Another example where high-dimensional covariates interact through an underlying network structure arises in spatial transcriptomics. This is a relatively new direction in biology made possible by technologies for massively parallelized measurement of cell transcriptomes/proteomes in situ that, unlike standard single cell sequencing methods, retains information regarding the spatial neighborhood of the cells. The spatial context of cells can be encoded into a nearest-neighbor graph or a Voronoi neighborhood graph/Delaunay triangulation (de Berg et al., 2008), where the nodes are cells and edges link cells that are situated proximal to each other (Eng et al., 2019; Goltsev et al., 2018; Palla et al., 2022; Perkel, 2019). Each node has a high dimensional feature set, encoding the measurements made for that cell, be it gene expression or protein expression, depending on the experimental protocol. Then, the goal is to understand how the spatial niche of a cell contributes to its phenotype (see Section 5 for more details). For other examples of network peer-effect in the presence of covariates, see Bertrand et al. (2000); Christakis and Fowler (2013); Duflo and Saez (2003); Glaeser et al. (1996); Sacerdote (2001); Trogdon et al. (2008) and references therein.

In the aforementioned examples, it is natural to envisage a model that combines the logistic model in (1) (which encodes the node-specific covariates) and the Ising model (2) (for capturing the network dependency). Towards this, Daskalakis et al. (2020) recently proposed the following model: Suppose for each node $1 \leqslant i \leqslant N$ of a network G_N on N vertices, one observes a d-dimensional covariate $\boldsymbol{Z}_i \in \mathbb{R}^d$. Then, the joint distribution of the binary outcomes $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$, conditioned on the network G_N and the observed covariates $\boldsymbol{Z} := (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N)^\top$ is given by:

$$\mathbb{P}(\boldsymbol{X}\big|\boldsymbol{Z}) \propto \exp\left(\frac{\beta}{2}\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} + \sum_{i=1}^{N} X_i(\boldsymbol{\theta}^\top \boldsymbol{Z}_i)\right) \tag{3}$$

where $\boldsymbol{A} = ((a_{ij}))_{1 \leqslant i,j \leqslant N}$ is the (appropriately scaled) adjacency matrix of the network G_N, the parameter β is a measure of the strength of dependence (the network 'peer effect'), and the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ measures the individual effects of the d-covariates. Specifically, as β becomes more positive, the outcomes of the nodes tend to align with those of their neighbors. On the other hand, when β is negative (which is also allowed in our theoretical framework), every node receives negative influences from its neighbors, a phenomenon referred to as *antiferromagnetism* in the statistical physics literature. Note that as in the classical Ising model (2), the quadratic term captures the overall dependency in the network arising from pairwise interactions, while the linear term $\boldsymbol{\theta}^\top \boldsymbol{Z}_i$ encodes the strength of the covariates on the outcome at the i-th node, for $1 \leqslant i \leqslant N$, as in the logistic model (1). Moreover, when $\beta = 0$, which corresponds to the independent case, (3) reduces to (1); hence, (3) can be viewed as a model for logistic regression with dependent observations.

The increasing relevance of models (2) and (3) for understanding covariate effects and nearest-neighbor interactions in network data, has made it imperative to develop computationally tractable methods for estimating the model parameters and understanding the statistical properties (rates of convergence) of the resulting estimates. A typical problem of interest is estimating the parameters of the model given a single sample of binary outcomes from an underlying network. For the classical Ising model as in (2), this problem has been extensively studied, beginning with the classical results on consistency and optimality of the maximum likelihood (ML) estimates for models where the underlying network is a spatial lattice (Comets, 1992; Gidas, 1988; Guyon and Künsch, 1992; Pickard, 1987). However, for general networks, parameter estimation using the ML method turns out to be notoriously hard due to the presence of an intractable normalizing constant in the likelihood. One approach to circumvent this issue that has turned out to be particularly useful, is the *maximum pseudolikelihood* (MPL) estimator (Besag, 1974, 1975). This provides a computationally efficient method for estimating the parameters of a Markov random field that avoids computing the normalizing constant, by maximizing an approximation to the likelihood function (a 'pseudo-likelihood'), based on conditional distributions. This approach was originally explored in the seminal paper of Chatterjee (2007), where general conditions for $\sqrt{N}$-consistency of the MPL estimate for the model (2) were derived.[2] This result was subsequently extended to more general models in Bhattacharya and Mukherjee (2018); Daskalakis et al. (2020, 2019b); Ghosal and Mukherjee (2020); Mukherjee et al. (2018); Mukherjee and Ray (2022); Mukherjee et al. (2022). In particular, for model (3) Daskalakis et al. (2019b) showed that given a single sample of observations $(X_i, \boldsymbol{Z}_i)_{1 \leqslant i \leqslant N}$ from (3), the MPL estimate of the parameters $(\beta, \boldsymbol{\theta})$ is $\sqrt{N}$-consistent, when the dimension d is *fixed*, under various regularity assumptions on the underlying network and the parameters. This result has been subsequently extended to models with higher-order interactions in Daskalakis et al. (2019b). The high-dimensional analogue of this problem under an ℓ_1 norm constraint on the regression parameter has been studied very recently in Kandiros et al. (2021).

In this paper, we consider the problem of parameter estimation in model (3) in the *high-dimensional regime* with sparsity constraints, that is, the number of covariates d grows with the size of the network N, but there are only a few non-zero regression coefficients. In other

2. A sequence of estimators $\{\hat{\beta}_N\}_{N \geqslant 1}$ is said to be $\sqrt{N}$-*consistent* at β, if for every $\delta > 0$, there exists $M := M(\delta, \beta) > 0$ such that $\mathbb{P}(\sqrt{N}|\hat{\beta}_N(\boldsymbol{X}) - \beta| \leqslant M) > 1 - \delta$, for all N large enough.

words, we assume that the parameter vector $\boldsymbol{\theta}$ is s-sparse, that is $||\boldsymbol{\theta}||_0 := \sum_{i=1}^d \mathbf{1}\{\theta_i \neq 0\} \leqslant s$, because, despite the fact that many covariates are available, we only expect a few of them to be relevant. Under this assumption, we want to estimate the parameters $(\beta, \boldsymbol{\theta})$ given a *single* sample of dependent observations $(X_i, \boldsymbol{Z}_i)_{1 \leqslant i \leqslant N}$ from model (3). One of the main reasons this problem is especially delicate, is that we only have access to a *single (dependent) sample* from the underlying model. Consequently, classical results from the M-estimation framework, which require *multiple independent samples*, are inapplicable. To deal with this dependence (which leads to the intractable normalizing constant as mentioned before) and the high-dimensionality of the parameter space, we propose a penalized maximum pseudo-likelihood approach for estimating the parameters. To this end, note that the conditional distribution of X_i given $(X_j)_{j \neq i}$ is:

$$\mathbb{P}(X_i | (X_j)_{j \neq i}, \boldsymbol{Z}) = \frac{e^{X_i \boldsymbol{\theta}^\top \boldsymbol{Z}_i + \beta X_i m_i(\boldsymbol{X})}}{e^{\boldsymbol{\theta}^\top \boldsymbol{Z}_i + \beta m_i(\boldsymbol{X})} + e^{-\boldsymbol{\theta}^\top \boldsymbol{Z}_i - \beta m_i(\boldsymbol{X})}}, \tag{4}$$

where, as before, $m_i(\boldsymbol{X}) = \sum_{j=1}^N a_{ij} X_j$ is the *local-effect* at node i, for $1 \leqslant i \leqslant N$. Therefore, by multiplying (4) over $1 \leqslant i \leqslant N$ and taking logarithms, we get the (negative) *log-pseudolikelihood* (LPL) function

$$\begin{aligned} L_N(\beta, \boldsymbol{\theta}) &= -\frac{1}{N} \sum_{i=1}^N \log \mathbb{P}(X_i \big| (X_j)_{j \neq i}, \boldsymbol{Z}) \\ &= -\frac{1}{N} \sum_{i=1}^N \{X_i(\boldsymbol{\theta}^\top \boldsymbol{Z}_i + \beta m_i(\boldsymbol{X})) - \log \cosh(\boldsymbol{\theta}^\top \boldsymbol{Z}_i + \beta m_i(\boldsymbol{X}))\} + \log 2. \end{aligned} \tag{5}$$

To encode the sparsity of the high-dimensional parameters, we propose a *penalized maximum pseudo-likelihood* (PMPL) estimator of $(\beta, \boldsymbol{\theta}^\top)$, which, given a regularization (tuning) parameter λ, is defined as:

$$(\hat{\beta}, \hat{\boldsymbol{\theta}}^\top) := \arg\min_{(\beta, \boldsymbol{\theta})} \{L_N(\beta, \boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_1\}, \tag{6}$$

where $||\boldsymbol{\theta}||_1 := \sum_{i=1}^d |\theta_i|$. Under various regularity assumptions, we prove that if λ is chosen proportional to $\sqrt{\log d / n}$, then with high probability,

$$\|(\hat{\beta}, \hat{\boldsymbol{\theta}}^\top) - (\beta, \boldsymbol{\theta}^\top)\| \lesssim_s \sqrt{\log d / N}, \tag{7}$$

whenever $d \to \infty$ such that $d = o(N)$ (Theorem 1). In particular, it follows from our results that for bounded degree networks, the PMPL estimate attains the same rate as in the independent logistic case (1), when $d = o(N)$ and the sparsity is bounded. We also have a more general result that quantifies the dependence on the network sparsity in the rate (7), which allows us to establish consistency of the PMPL estimate beyond bounded degree graphs (Proposition 7). Our results are fundamentally different from existing results on parameter estimation in high-dimensional graphical models based on multiple i.i.d. samples (see Section 1.1 for a review). Here, we only have access to a single sample from the model (3), hence, unlike in the multiple samples case, one cannot treat the different neighborhoods in the network as independent, which renders classical techniques for proving consistency

inapplicable. Consequently, to handle the dependencies among the different neighborhoods in the pseudo-likelihood function we need to use a different (non-classical) set of tools. Specifically, our proofs combine a conditioning technique introduced in Dagan et al. (2021), which tranforms a general Ising model to a model satisfying the Dobrushin condition (where the dependence is sufficiently weak), and the concentration inequalities for functions of dependent random variables in the Dobrushin regime, based on the method of exchangeable pairs (Chatterjee, 2016).

Next, we study the effect of dependence on estimating the regression parameters $\boldsymbol{\theta}$. Specifically, we want to understand how the presence of dependence through the underlying network structure effects the rate of estimation of the high-dimensional regression coefficient under sparsity constraints. While there have been several recent attempts to understand the implications of dependence in high-dimensional inference, most of them have focused on Gaussian models. Going beyond Gaussian models, Mukherjee, Mukherjee, and Yuan (2018) and Deb et al. (2020) considered the problem of testing the global null hypothesis (that is, $\boldsymbol{\theta} = \mathbf{0}$) against sparse alternatives in a special case of model (3) (where $d = N$ and the design matrix $\boldsymbol{Z} = \boldsymbol{I}_N$ is the identity). However, to the best of our knowledge, the effect of dependence on parameter estimation in Ising models with covariates has not been explored before. In the sequel, we establish that the PMPL estimate for $\boldsymbol{\theta}$ in model (3), given a dependence strength β and the sparsity constraint $\|\boldsymbol{\theta}\|_0 = s$, attains the classical $O(\sqrt{s \log d/N})$ rate, *despite the presence of dependence*, in the entire high-dimensional regime (where d can be much larger than N) and also recovers the correct dependence on the sparsity s (see Theorem 2). As a consequence, we establish that the PMPL estimate is $O(1/\sqrt{N})$-consistent (up to logarithmic factors) for the model (3) in various natural sparse graph ensembles, such as Erdős-Rényi random graphs and inhomogeneous random graphs that include the popular stochastic block model (Theorem 3 and Corollary 5). We also develop a proximal gradient algorithm for efficiently computing the PMPL estimate and evaluate its performance in numerical experiments for Erdős-Rényi random graphs, inhomogeneous random graph models, such as the stochastic block model and the β-model, and the preferential attachment model (Section 4). Finally, in Section 5, we illustrate the effectiveness of the proposed model in selecting relevant genes for clustering spatial gene expression data.

1.1 Related Work

The asymptotic properties of penalized likelihood methods for logistic regression and generalized linear models in high dimensional settings have been extensively studied (see, for example, Bach (2010); Bunea (2008); Kakade et al. (2010); Meier et al. (2008); Salehi et al. (2019); van de Geer (2008); van de Geer et al. (2014) and references therein). These results allow d to be much bigger than N and provide rates of convergence for the high-dimensional parameters under various sparsity constraints. The performance of the ML estimate in the logistic regression model when the dimension d scales proportionally with N has also been studied in a series of recent papers (Candès and Sur, 2020; Sur and Candès, 2019; Sur et al., 2019). However, as discussed earlier, ML estimation is both computationally and mathematically intractable in model (3), because of the dependency induced by the underlying network. That we are able to recover rates similar to those in the classical high-dimensional

logistic regression, in spite of this underlying dependency, using the PMPL method, is one of the highlights of our results.

The problem of estimation and structure learning in graphical models and Markov random fields is another related area of active research. Here, the goal is to estimate the model parameters or learn the underlying graph structure given access to *multiple* i.i.d. samples from a graphical model. For more on these results refer to Anandkumar et al. (2012); Bresler (2015); Bresler and Karzand (2020); Chow and Liu (1968); Guo et al. (2011); Hamilton et al. (2017); Klivans and Meka (2017); Ravikumar et al. (2010); Santhanam and Wainwright (2012); Vuffray et al. (2016) and references therein. In particular, Ravikumar et al. (2010) and Xue et al. (2012) use a regularized pseudo-likelihood approach to learn the structure of the interaction $\boldsymbol{A}$ given *multiple* i.i.d. samples from an Ising model. In a related direction, Daskalakis et al. (2019a) studied the related problems of identity and independence testing, and Cao et al. (2022); Neykov and Liu (2019) considered problems in graph property testing, given access to multiple samples from an Ising model.

All the aforementioned results, however, are in contrast with the present work, where the underlying graph structure is assumed to be *known* and the goal is to derive rates of estimation for the parameters given a *single* sample from the model. This is motivated by the applications mentioned above where it is often difficult, if not impossible, to generate many independent samples from the model within a reasonable amount of time. More closely related to the present work are results of Li and Zhang (2010) and Li et al. (2015) on Bayesian methods for variable selection in high-dimensional covariate spaces with an underlying network structure, where an Ising prior is used on the model space for incorporating the structural information. Recently, Kim et al. (2021) developed a variational Bayes procedure using the pseudo-likelihood for estimation based on a single sample in a two-parameter Ising model. Convergence of coordinate ascent algorithms for mean field variational inference in the Ising model has been recently analyzed in Plummer et al. (2020).

For continuous response with an underlying network/spatial dependence structure, a related model is the well-known spatial autoregressive (SAR) model and its variants, where likelihood estimation based on conditional distributions have been used as well (see Huang et al. (2019); Lee (2004); Lee et al. (2010); Zhu et al. (2020) and the references therein).

1.2 Notation

The following notation will be used throughout the remainder of the paper. For a vector $\boldsymbol{a} := (a_1, \ldots, a_s) \in \mathbb{R}^s$ and $0 < p < \infty$, $\|\boldsymbol{a}\|_p := (\sum_{i=1}^s |a_i|^p)^{\frac{1}{p}}$ denotes its p-th norm and $\|\boldsymbol{a}\|_\infty := \max_{1 \leqslant i \leqslant s} |a_s|$ its maximum norm, respectively. Moreover, $\|\boldsymbol{a}\|_0 := \sum_{i=1}^s \mathbf{1}\{a_i \neq 0\}$ denote the 'zero-norm' of $\boldsymbol{a}$, which counts the number of non-zero coordinates of $\boldsymbol{a}$.

For a matrix $\boldsymbol{M} := ((M_{ij}))_{1 \leqslant i \leqslant s, 1 \leqslant j \leqslant t} \in \mathbb{R}^{s \times t}$ we define the following norms:

- $\|\boldsymbol{M}\|_F := \sqrt{\sum_{i=1}^s \sum_{j=1}^t M_{ij}^2}$,
- $\|\boldsymbol{M}\|_\infty := \max_{1 \leqslant i \leqslant s} \sum_{j=1}^t |M_{ij}|$,
- $\|\boldsymbol{M}\|_1 := \max_{1 \leqslant j \leqslant t} \sum_{i=1}^s |M_{ij}|$,
- $\|\boldsymbol{M}\|_2 := \sigma_{\max}(\boldsymbol{M})$, where $\sigma_{\max}(\boldsymbol{M})$ denotes the largest singular value of $\boldsymbol{M}$.

Note that if $\boldsymbol{M}$ is a square matrix, then $\|\boldsymbol{M}\|_2 = \max\{|\lambda_{\min}(\boldsymbol{M})|, |\lambda_{\max}(\boldsymbol{M})|\}$, where $\lambda_{\max}(\boldsymbol{M})$ and $\lambda_{\min}(\boldsymbol{M})$ denote the maximum and minimum eigenvalues of $\boldsymbol{M}$, respectively.

For positive sequences $\{a_n\}_{n\geqslant 1}$ and $\{b_n\}_{n\geqslant 1}$, $a_n = O(b_n)$ means $a_n \leqslant C_1 b_n$, $a_n = \Omega(b_n)$ means $a_n \geqslant C_2 b_n$, and $a_n = \Theta(b_n)$ means $C_2 n \leqslant a_n \leqslant C_1 b_n$, for all n large enough and positive constants C_1, C_2. Similarly, $a_n \lesssim b_n$ means $a_n = O(b_n)$, $a_n \gtrsim b_n$ means $a_n = \Omega(b_n)$. Moreover, subscripts in the above notations, for example $\lesssim_{\square}$, $O_{\square}$ and $\Omega_{\square}$ denote that the hidden constants may depend on the subscripted parameters $\square$. Finally, we say that $a_n = \tilde{O}(b_n)$, if $a_n \leqslant C_1(\log n)^{r_1} b_n$ and $a_n = \tilde{\Theta}(b_n)$, if $C_2(\log n)^{r_2} b_n \leqslant a_n \leqslant C_1(\log n)^{r_1} b_n$, for all n large enough and some positive constants C_1, C_2, r_1, r_2.

1.3 Organization

The remainder of the paper is organized as follows. The rates of consistency of the estimates are presented in Section 2. In Section 3, we apply our results to various common network models. The algorithm for computing the estimates and simulation results are presented in Section 4. The proofs of the technical results are given in the Appendix.

2. Rates of Consistency

Next, we present our results on rates of convergence of the PMPL estimator. In Section 2.1, we present the rates of convergence of the PMPL estimates $(\hat{\beta}, \hat{\boldsymbol{\theta}}^\top)$. The rate for estimating the regression parameters is presented in Section 2.2.

2.1 Consistency of the PMPL Estimate

We begin by stating the relevant assumptions:

Assumption 1 *The interaction matrix $\boldsymbol{A}$ in (3) satisfies the following comdition:*

$$\sup_{N\geqslant 1} \|\boldsymbol{A}\|_\infty < \infty.$$

Assumption 2 *The design matrix $\boldsymbol{Z} := (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)^\top$ satisfies*

$$\liminf_{N\to\infty} \lambda_{\min}\left(\frac{1}{N}\boldsymbol{Z}^\top \boldsymbol{Z}\right) > 0.$$

Assumption 3 *The signal parameters $\boldsymbol{\theta}$ and the covariates $\{\boldsymbol{Z}_i\}_{1\leqslant i\leqslant N}$ are uniformly bounded, that is, there exist positive constants Θ and M such that $\|\boldsymbol{\theta}\|_\infty < \Theta$ and $\|\boldsymbol{Z}_i\|_\infty < M$, for all $1 \leqslant i \leqslant N$.*

Under the above assumptions we establish the rate of convergence of the PMPL estimate (6) given a single sample of observations from the model (3), when the parameter vector is sparse, that is, $\|(\beta, \boldsymbol{\theta}^\top)^\top\|_0 = s$. For notational convenience, we henceforth denote the $(d+1)$-dimensional vector of parameters by $\boldsymbol{\gamma} := (\beta, \boldsymbol{\theta}^\top)^\top$ and the $(d+1)$-dimensional vector of PMPL estimates obtained from (6) by $\hat{\boldsymbol{\gamma}} = (\hat{\beta}, \hat{\boldsymbol{\theta}}^\top)^\top$.

Theorem 1 *Suppose that Assumptions 1, 2, 3 hold, and* $\liminf_{N\to\infty} \frac{1}{N}\|\boldsymbol{A}\|_F^2 > 0$. *Then, there exists a constant* $\delta > 0$ *such that by choosing* $\lambda := \delta\sqrt{\log(d+1)/N}$ *in the objective function in* (22) *we have,*

$$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2 = O_s\left(\sqrt{\frac{\log d}{N}}\right) \quad \text{and} \quad \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 = O_s\left(\sqrt{\frac{\log d}{N}}\right), \tag{8}$$

with probability $1 - o(1)$, *as* $N \to \infty$ *and* $d \to \infty$ *such that* $d = o(N)$.

The conditions in Theorem 1 combined aim to strike a balance between the signal from the peer-effects to that from the covariates, to ensure consistent estimation for all β. In particular, a control on $\|\boldsymbol{A}\|_\infty$ is required to ensure that the peer effects coming from the quadratic dependence term in the probability mass function (3) do not overpower the effect of the signal $\boldsymbol{\theta}$ coming from the linear terms $\boldsymbol{\theta}^\top \boldsymbol{Z}_i$, thereby hindering joint recovery of the correlation term β and the signal term $\boldsymbol{\theta}$. At the same time, we also require the interaction matrix to be not too sparse, and its entries to be not too small, in order to ensure that the effect of the correlation parameter β is not nullified. This is guaranteed by the condition $\|\boldsymbol{A}\|_F^2 = \Omega(N)$. For example, when $\boldsymbol{A}$ is the scaled adjacency of a graph G_N, then Assumption 1 together with the condition $\|\boldsymbol{A}\|_F^2 = \Omega(N)$ implies that G_N has bounded maximum degree (see Section 3). In fact, in the proof we keep track of the dependence on $\|\boldsymbol{A}\|_F$ in the error rate (see Proposition 7 in Section A), which allows us to establish consistency of the PMPL estimate beyond bounded degree graphs (see Section 3.3 for details).

Remark 1 It is worth noting that it may be impossible to estimate $(\beta, \boldsymbol{\theta})$ consistently without any diverging lower bound on $\|\boldsymbol{A}\|_F^2 = \Omega(N)$ or, in other words, if the graph G_N is too dense. This phenomenon is observed in the Curie-Weiss Model (where the interaction matrix $a_{ij} = 1/N$, for $1 \leqslant i \neq j \leqslant N$) (Comets and Gidas, 1991). In this example, each entry of $\boldsymbol{A}$ is $O(1/N)$, and hence, $\|\boldsymbol{A}\|_F^2 = O(1)$, and even when $d = 1$ (and $\boldsymbol{Z}_1 = \ldots = \boldsymbol{Z}_N$) consistent estimation of the parameters β and θ is impossible (see Theorem 1.13 in Ghosal and Mukherjee (2020)).

The proof of Theorem 1 is given in Section A. As mentioned before, this is a consequence of a more general result which gives rates of convergence for the PMPL estimate in terms of the $\|\boldsymbol{A}\|_F$ (Proposition 7). Broadly speaking, the proof involves the following two steps:

- *Concentration of the gradient*: The first step in the proof of Theorem 1 is to show that the gradient of the logarithm of the pseudo-likelihood function L_N (recall (5)) is concentrated around zero in the ℓ_∞ norm. For this step, we use the conditioning trick introduced in Dagan et al. (2021), which reduces a general Ising model to an Ising model in the high-temperature regime, where exponential concentration inequalities for functionals of Ising models are available (Chatterjee, 2016). The details are formalized in Lemma 24.

- *Strong-concavity of the pseudo-likelihood*: In the second step, we show that the logarithm of the pseudo-likelihood function is strongly concave with high probability.

This entails showing that the lowest eigenvalue of the Hessian of L_N is bounded away from zero with high probability. Towards this, the minimum eigenvalue condition in Assumption 2, which is standard in the high-dimensional literature (see Loh (2017); Ravikumar et al. (2010)), is crucial. In particular, this condition holds with high probability, if the covariates $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ are i.i.d. realizations from a sub-Gaussian distribution on $\mathbb{R}^d$ (see Theorem 2.1 in Daskalakis et al. (2019b)). Under Assumption 2 and using lower bounds on the variance of linear projections of $\boldsymbol{X}$ developed in Dagan et al. (2021) and concentration results from Adamczak et al. (2019), we establish the strong-concavity of the pseudo-likelihood in Lemma 9.

Remark 2 Note that the rate in (8) suppresses the dependence on the sparsity parameter s in the order term. From the proof of Theorem 1, it will be seen that the dependence is, in general, exponential in s. However, if one replaces Assumption 3 with the stronger assumption that the ℓ_2 norms of the parameters and the covariates are *bounded*, that is, $\|\boldsymbol{\theta}\|_2 < \Theta$ and $\|\boldsymbol{Z}_i\|_2 < M$, for all $1 \leqslant i \leqslant N$, then our proof can be easily modified to recover the standard high-dimensional $O(\sqrt{s \log d/N})$ rate (see Remark 16). In fact, this stronger assumption has been used recently in Kandiros et al. (2021) to derive rates of the pseudo-likelihood estimate under ℓ_1 sparsity. Specifically, (Kandiros et al., 2021, Theorem 2) showed that if $\|\boldsymbol{\gamma}\|_1 \leqslant s$ and the parameters and the covariates are ℓ_2-bounded, their estimate $\tilde{\boldsymbol{\gamma}}$ under Assumption 1 satisfies:

$$\|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2 = O\left(\left(\frac{s \log d}{N}\right)^{\frac{1}{6}}\right),$$

with high probability. Note that the dependence on N in the RHS above is worse than the expected $1/\sqrt{N}$-rate. Moreover, the ℓ_2-boundedness of the covariates is quite restrictive in the high-dimensional setup. On the other hand, this work derives rates under ℓ_0 sparsity and a more realistic ℓ_∞-bounded condition (Assumption 3). Under this condition, we are able to derive the correct dependence on N and d (and also on s, if the stronger ℓ_2-boundedness is imposed as mentioned above) in the regime where $d = o(N)$. An exponential dependence on d also appears in Daskalakis et al. (2019b) (see footnote in page 4), where the convergence rate of the MPL estimate is derived in the fixed d regime. This rate can be improved to $O(\sqrt{d/N})$ under the ℓ_2- boundedness assumption (see, for example, Daskalakis et al. (2020)). Our results show that this can be further improved to $O(\sqrt{\log d/N})$ in the regime of constant sparsity.

2.2 Estimation of the Regression Coefficients

In this section, we consider the problem of estimating the regression coefficients $\boldsymbol{\theta}$, for fixed β. The goal is to understand how network dependence may affect our ability to estimate the high-dimensional regression coefficient under sparsity constraints. Towards this, we study the properties of the following PMPL estimator for the regression coefficients $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}} \{L_{\beta,N}(\boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_1\}, \tag{9}$$

where λ is a regularization parameter and (recalling (5))

$$L_{\beta,N}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\{X_i(\boldsymbol{\theta}^\top \boldsymbol{Z}_i + \beta m_i(\boldsymbol{X})) - \log\cosh(\boldsymbol{\theta}^\top \boldsymbol{Z}_i + \beta m_i(\boldsymbol{X}))\} + \log 2. \tag{10}$$

To handle the high-dimensional regime, we need to make the following assumption on the design matrix $\boldsymbol{Z} := (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)^\top$. To this end, we define the Rademacher complexity of the $\{\boldsymbol{Z}_i\}_{1\leqslant i\leqslant N}$ as:

$$\mathcal{R}_N := \mathbb{E}\left(\left\|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i \boldsymbol{Z}_i\right\|_\infty\right), \tag{11}$$

where $\{\varepsilon_i\}_{1\leqslant i\leqslant N}$ is a sequence of i.i.d. Rademacher random variables and the expectation in (11) is taken jointly over the randomness of $\{\boldsymbol{Z}_i\}_{1\leqslant i\leqslant N}$ and $\{\varepsilon_i\}_{1\leqslant i\leqslant N}$.

Assumption 4 *Suppose the covariates $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N$ are drawn i.i.d. from a distribution with mean zero and satisfying the following conditions:*

(1) *There exist positive constants κ_1, κ_2 such that*

$$\mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}_1\rangle^2\right) \geqslant \kappa_1 \text{ and } \mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}_1\rangle^4\right) \leqslant \kappa_2,$$

for all $\boldsymbol{\eta} \in \mathbb{R}^d$ such that $\|\boldsymbol{\eta}\|_2 = 1$.

(2) $\mathcal{R}_N = O(\sqrt{\log d/N})$.

(3) *There exists a constant $C > 0$ such that $\max_{1\leqslant j\leqslant d}\frac{1}{N}\sum_{i=1}^{N} Z_{ij}^2 \leqslant C$ holds with probability 1.*

These types of conditions are standard in the high-dimensional statistics literature (see Bickel et al. (2009); Candes and Tao (2007); Meinshausen and Yu (2009); Negahban et al. (2012); Raskutti et al. (2011); van de Geer and Bühlmann (2009) and references therein), which are known to hold for many natural classes of design matrices. For example, if $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n$ are i.i.d. sub-Gaussian random variables with mean zero, then Wainwright (2019, Exercise 9.8) implies that $\mathcal{R}_N = O(\sqrt{\log d/N})$.

Remark 3 As mentioned before, when $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n$ are i.i.d Gaussian with mean zero and covariance matrix Σ, then Assumption 4 (2) holds (by Wainwright (2019, Exercise 9.8)). To understand when Assumption 4 (1) holds, note that for $\boldsymbol{\eta} \in \mathbb{R}^d$ such that $\|\boldsymbol{\eta}\|_2 = 1$ we have

$$\mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^2\right) = \mathbb{E}(N(0, \boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta})^2) = \boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta} \geqslant \lambda_{\min}(\Sigma)$$

and

$$\mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^4\right) = \mathbb{E}(N(0, \boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta})^4) = 3(\boldsymbol{\eta}^\top\Sigma\boldsymbol{\eta})^4 \leqslant 3\lambda_{\max}^2(\Sigma),$$

where $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ are the minimum and maximum eigenvalues of Σ, respectively. Therefore, Assumption 4 (1) holds, if we assume that there exist positive constants c_*, c^*, such that the covariance matrix Σ satisfies $c_* \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant c^*$.

Under the above assumptions, we now show in the following theorem that one can consistently estimate the regression parameters of the model (3) at the same rate as the classical (independent) logistic regression model (1).

Theorem 2 *Fix $\beta \in \mathbb{R}$. Suppose the interaction matrix $\boldsymbol{A}$ in (3) satisfies Assumptions 1 and the covariates $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N$ satisfy Assumption 4. Moreover, assume that there exists a positive constant Θ such that $\|\boldsymbol{\theta}\|_2 \leqslant \Theta$. Then, there exists a constant $\delta > 0$ such that by choosing $\lambda := \delta\sqrt{\log d/N}$ in the objective function in (9) we have,*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = O\left(\sqrt{\frac{s \log d}{N}}\right) \quad \text{and} \quad \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 = O\left(s\sqrt{\frac{\log d}{N}}\right),$$

with probability $1 - o(1)$, as $N, d \to \infty$ such that $s\sqrt{\log d/N} = o(1)$.

The proof of Theorem 2 is given in Section B in the Appendix. We follow the strategy outlined in Wainwright (2019, Chapter 9) for showing rates of consistency for high-dimensional generalized linear models. In particular, we show that the pseudo-likelihood loss function satisfies the restricted strong concavity condition (Proposition 19) under Assumption 4. Consequently, we can establish the consistency of the PMPL estimate of the regression parameters in the entire high-dimensional regime (where d can be much larger than N) and also recover the correct dependence on the sparsity s.

Remark 4 Note that, unlike in Theorem 1, the Frobenius norm assumption $\|\boldsymbol{A}\|_F^2 = \Omega(N)$, is not required in Theorem 2. In particular, the only assumption on $\boldsymbol{A}$ one needs in Theorem 2 is $\|\boldsymbol{A}\|_\infty < 1$ (Assumption 1). For example, when $\boldsymbol{A}$ is the scaled adjacency matrix of a graph G_N, the assumption $\|\boldsymbol{A}\|_\infty < 1$ is equivalent to the maximum degree of G_N being of the same order as the average degree of G_N (see (13)). This is expected because when β is known, the parameter $\boldsymbol{\theta}$ can be estimated at the classical high-dimensional rate, irrespective of the total edge density of the network, as long as the peer effects coming from the quadratic dependence term do not overshadow effect of the linear term $\boldsymbol{\theta}^\top \boldsymbol{Z}$, which is ensured by the condition $\|\boldsymbol{A}\|_\infty < 1$. This condition, in particular, implies that no node of the network has an unduly large effect on the corresponding model, and is satisfied by most Ising models that are commonly studied in the literature.

3. Application to Various Network Structures

In this section, we apply Theorem 1 to establish the consistency of the PMPL estimator (22) for various natural network models. To this end, let $G_N = (V(G_N), E(G_N))$ be a sequence of graphs with $V(G_N) = [N] := \{1, 2, \ldots, N\}$ and adjacency matrix $\mathcal{A}(G_N) = ((a_{ij}))_{1 \leqslant i,j \leqslant N}$. We denote by d_v the degree of the vertex $v \in V(G_n)$. To ensure that the model (3) has non-trivial scaling properties, one needs to chose the interaction matrix $\boldsymbol{A}$ as the scaled adjacency matrix of G_N. In particular, define

$$\boldsymbol{A}_{G_N} := \frac{N}{2|E(G_N)|} \cdot \mathcal{A}(G_N) \tag{12}$$

Throughout this section, we consider model (3) with $\boldsymbol{A} = \boldsymbol{A}_{G_N}$ as above. We also assume that the number of non-isolated vertices in G_N (that is, the number of vertices in G_N with

degree at least 1) is $\Omega(N)$. Note that this implies, $|E(G_N)| \gtrsim N$. Finally, we also assume that the sparsity $s = O(1)$ and consequently, absorb the dependence on s in the O-terms (recall Remark 2).

3.1 Bounded Degree Graphs

A sequence of graphs $\{G_N\}_{N\geqslant 1}$ is said to have bounded maximum degree if its maximum degree is uniformly bounded, that is, $\sup_{N\geqslant 1} d_{\max} < \infty$, where $d_{\max} := \max_{v\in V(G_n)} d_v$ is the maximum degree of G_N. Note that if G_N has bounded maximum degree and has $\Omega(N)$ non-isolated vertices, then $|E(G_N)| = \Theta(N)$.

Networks arising in certain applications, especially those with an underlying spatial or lattice structure generally have bounded degree. These include planar maps which encode neighborhood relations (Batra et al., 2010; Johnson et al., 2016), lattice models for capturing nearest-neighbor interactions between image pixels, and demand-aware networks (Avin et al., 2020) among others. It is easy to check that the conditions in Assumption 1 are satisfied for bounded degree graphs. Towards this, note that, under the scaling in (12), the condition $\sup_{N\geqslant 1} \|\boldsymbol{A}\|_\infty < \infty$ is equivalent to $\|\boldsymbol{A}_{G_N}\|_\infty = \frac{N}{2|E(G_N)|} \max_{v\in V(G_n)} d_v$. Hence, under the scaling in (12), the assumption $\sup_{N\geqslant 1} \|\boldsymbol{A}_{G_N}\|_\infty < \infty$ is equivalent to

$$d_{\max} := \max_{v\in V(G_n)} d_v = O\left(\frac{|E(G)|}{N}\right), \tag{13}$$

that is, the maximum degree of G_N is of the same order as its average degree. Moreover, the condition $\liminf_{N\to\infty} \frac{1}{N}\|\boldsymbol{A}_{G_N}\|_F^2 > 0$ is equivalent to

$$\limsup_{N\to\infty} \frac{|E(G_N)|}{N} < \infty, \tag{14}$$

that is, the average degree of G_N is bounded. Therefore, (13) and (14) are together equivalent to the condition that G_N has bounded maximum degree. Hence, the PMPL estimate is $\sqrt{\log d/N}$-consistent for any sequence of graphs of bounded maximum degree, whenever the assumptions in Theorem 1 hold.

3.2 Sparse Inhomogeneous Random Graphs

Although Theorem 1 requires that the maximum degree of G_N has to be of the same order as the average degree (see (13)), our proofs can be easily adapted to establish similar rates of consistency of the PMPL estimate (up to polylog(N) factors), if the maximum degree G_N grows poly-logarithmically with respect to the average degree, which, in particular, is the case for sparse inhomogeneous random graphs. This is summarized in the following result. The proof is given in Section C.1 of the Appendix.

Theorem 3 *Suppose $\{G_N\}_{N\geqslant 1}$ is a sequence of graphs with $|E(G_N)| = O(N)$, $d_{\max} = \tilde{O}(1)$, and $\Omega(N)$ non-isolated vertices. Then for $\lambda = \tilde{\Theta}\left(\sqrt{\log d/N}\right)$,*

$$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2 = \tilde{O}\left(\frac{1}{\sqrt{N}}\right)$$

with probability $1 - o(1)$ as $N, d \to \infty$ such that $d = o(N)$.

Theorem 3 can be applied to obtain rates of convergence in sparse inhomogeneous random graph models.

Definition 4 *(Bollobás et al., 2007) Given a symmetric matrix $\boldsymbol{P}^{(N)} = ((p_{uv})) \in [0,1]^{N\times N}$ with zeroes on the diagonal, the inhomogeneous random graph $\mathcal{G}(N, \boldsymbol{P}^{(N)})$ is the graph with vertex set $[N] := \{1, 2, \ldots, N\}$ where the edge (u, v) is present with probability p_{uv}, independent of the other edges, for every $1 \leqslant u < v \leqslant N$.*

The class of inhomogeneous random graph models defined above includes several popular network models, such as the Chung-Lu model (Chung and Lu, 2002), the β-model (Chatterjee et al., 2011), random dot product graphs (Young and Scheinerman, 2007; Tang et al., 2017), and stochastic block models (Holland et al., 1983; Lei, 2016). Next, we consider the sparse regime wherein

$$\max_{1\leqslant u,v\leqslant N} p_{uv} = O\left(\frac{1}{N}\right). \tag{15}$$

In this regime, the expected degree remains bounded, although the maximum degree can diverge at rate $O(\log N)$ (Benaych-Georges et al., 2019; Krivelevich and Sudakov, 2003). We will also assume that there exists $\varepsilon \in (0,1)$ and $\Omega(N)$ vertices $u \in G_N$, such that

$$\limsup_{N\to\infty} \prod_{v=1}^{N} (1 - p_{uv}) < \varepsilon. \tag{16}$$

This will ensure $\mathcal{G}(N, \boldsymbol{P}^{(N)})$ has $\Omega(N)$ non-isolated vertices. Under these assumptions we have the following result:

Corollary 5 *Suppose G_N is a realization of the inhomogeneous random graph $\mathcal{G}(N, \boldsymbol{P}^{(N)})$, where $\boldsymbol{P}^{(N)}$ satisfies the conditions in (15) and (16). Then for $\lambda = \tilde{\Theta}\left(\sqrt{\log(d+1)/N}\right)$,*

$$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2 = \tilde{O}\left(\frac{1}{\sqrt{N}}\right),$$

with probability $1 - o(1)$ as $N, d \to \infty$ such that $d = o(N)$.

Corollary 5 is proved in Section C.2 of the Appendix. In the following example, we illustrate how it can be applied to sparse stochastic block models, in particular, sparse Erdős-Rényi random graphs.

Example 1 (Sparse stochastic block models) Fix $K \geqslant 1$, a vector of community proportions $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_K) \in (0,1)^K$, such that $\sum_{j=1}^{K} \lambda_j = 1$, and a symmetric probability matrix $\boldsymbol{B} := ((b_{ij}))_{1\leqslant i,j\leqslant K}$, where $b_{ij} \in [0,1]$, for all $1 \leqslant i, j \leqslant K$ and $b_{ij} > 0$ for some $1 \leqslant i, j \leqslant K$. The (sparse) stochastic block model with proportion vector $\boldsymbol{\lambda}$ and probability matrix $\boldsymbol{B}$ is the inhomogeneous random graph $\mathcal{G}(N, \boldsymbol{P}^{(N)})$, with $\boldsymbol{P}^{(N)} = ((p_{uv}))_{1\leqslant u,v\leqslant N}$ where

$$p_{uv} = \frac{b_{ij}}{N} \quad \text{for} \quad (u,v) \in B_i \times B_j, \tag{17}$$

where $B_j := (N\sum_{i=1}^{j-1}\lambda_i, N\sum_{i=1}^{j}\lambda_i]\bigcap[N]$, for $j \in \{1,\dots,K\}$. In other words, the set of vertices is divided into K blocks (communities) $B_1, B_2, \dots, B_K$, such that the edge between vertices $u \in B_i$ and $v \in B_j$ occurs independently with probability b_{ij}/N. Clearly, in this case (15) holds. Next, to check that (16) holds, choose $1 \leqslant i, j \leqslant K$ such that $b_{ij} > 0$. Then for all $u \in B_i$,

$$\limsup_{N\to\infty}\prod_{v=1}^{N}(1-p_{uv}) \leqslant \limsup_{N\to\infty}\prod_{v\in B_j}\left(1-\frac{b_{ij}}{N}\right) = \exp(-\lambda_j b_{ij}) < 1,$$

which verifies (16), since $|B_j| = \Omega(N)$. Hence, by Corollary 5, the PMPL estimate (6) is $\tilde{O}(1/\sqrt{N})$-consistent in this example. As a consequence, the PMPL estimate is $\tilde{O}(1/\sqrt{N})$-consistent for sparse Erdős-Rényi random graphs $G(N, c/N)$, which corresponds to setting $b_{ij} = c$, for all $1 \leqslant i, j \leqslant K$ in (17).

3.3 Beyond Bounded Degree Graphs

We can also establish the consistency of the PMPL estimate beyond bounded degree graphs using Proposition 7, which provides error rates in terms of $\|\boldsymbol{A}\|_F$. To this end, note that when $\boldsymbol{A} = \boldsymbol{A}_{G_N}$ is the scaled adjacency matrix of G_N as in (12), then

$$\|\boldsymbol{A}\|_F = \Theta\left(\frac{N}{\sqrt{|E(G_N)|}}\right).$$

Hence, whenever (13) holds, Proposition 7 implies,

$$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2 = O_s\left(\sqrt{\frac{|E(G_N)|^2 \log d}{N^3}}\right)$$

with probability $1 - o(1)$, whenever $d = o(N^2/|E(G_N)|)$. This shows that the PMPL estimate is consistent whenever $|E(G_N)| = o(N^{3/2})$ (up to log-factors) and $d = o(N^2/|E(G_N)|)$. In particular, if G_N is Δ-regular (that is, all vertices have of G_N has degree Δ), then the rate of convergence becomes $O_s(\Delta\sqrt{\log d/N})$, if $d = o(N/\Delta)$.

4. Computation and Experiments

Next, we discuss an algorithm for computing the PMPL estimates (Section 4.1) and evaluate its performance in numerical experiments using synthetic data (Section 4.2).

4.1 Computation of the PMPL Estimates

A classical method developed for solving sparse estimation problems is the proximal descent algorithm (Friedman et al., 2010). We employ this algorithm to the optimization problem (9). To describe the algorithm, let

$$f(\boldsymbol{z}) := L_N(\boldsymbol{z}) + \lambda||\boldsymbol{z}||_1, \tag{18}$$

for $\boldsymbol{z} \in \mathbb{R}^{d+1}$, with $L_N(\cdot)$ as defined in (5). Also, for $t \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^{d+1}$ define

$$G_t(\boldsymbol{x}) = \frac{1}{t}\left(\boldsymbol{x} - \text{prox}_t(\boldsymbol{x} - t\nabla L_N(\boldsymbol{x}))\right),$$

where

$$\text{prox}_t(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}\in\mathbb{R}^{d+1}} \left\{\frac{1}{2t}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + \lambda||\boldsymbol{z}||_1\right\} = \left(x_i\left(1 - \frac{t\lambda}{|x_i|}\right)_+\right)_{1\leqslant i\leqslant d+1},$$

is minimized by the soft thresholding estimator. To chose the step size in the proximal descent algorithm, we employ a backtracking line search, which is commonly used in gradient-based as well as in lasso-type problems (Qin et al., 2013). To justify this we invoke the following result applied to the function f defined in (18):

Proposition 6 *(Vandenberghe, 2022) Fix $s \geqslant 1$ and a step size $t > 0$. Suppose at the s-th iteration the following bound holds:*

$$L_N(\boldsymbol{\gamma}^{(s)} - tG_t(\boldsymbol{\gamma}^{(s)})) \leqslant L_N(\boldsymbol{\gamma}^{(s)}) - t\nabla L_N(\boldsymbol{\gamma}^{(s)})^\top G_t(\boldsymbol{\gamma}^{(s)}) + \frac{t}{2}\|G_t(\boldsymbol{\gamma}^{(s)})\|_2^2. \tag{19}$$

Then for all $\boldsymbol{\gamma} \in \mathbb{R}^{d+1}$,

$$f(\boldsymbol{\gamma}^{(s)} - tG_t(\boldsymbol{\gamma}^{(s)})) \leqslant f(\boldsymbol{\gamma}) + G_t(\boldsymbol{\gamma}^{(s)})^\top(\boldsymbol{\gamma}^{(s)} - \boldsymbol{\gamma}) - \frac{t}{2}\|G_t(\boldsymbol{\gamma}^{(s)})\|_2^2. \tag{20}$$

Note that setting $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(s)}$ in (20) gives,

$$f(\boldsymbol{\gamma}^{(s)} - tG_t(\boldsymbol{\gamma}^{(s)})) \leqslant f(\boldsymbol{\gamma}^{(s)}) - \frac{t}{2}\|G_t(\boldsymbol{\gamma}^{(s)})\|_2^2.$$

This shows that whenever the line-search condition (19) holds, the descent of the objective function is guaranteed by setting

$$\boldsymbol{\gamma}^{(s+1)} \leftarrow \boldsymbol{\gamma}^{(s)} - tG_t(\boldsymbol{\gamma}^{(s)}) = \text{prox}_t(\boldsymbol{\gamma}^{(s)} - t\nabla L_N(\boldsymbol{\gamma}^{(s)})).$$

Therefore, the proximal gradient descent algorithm for optimization problem (9), with step size chosen by backtracking line search, proceeds in the following two steps: We initialize with $\boldsymbol{\gamma}^{(0)} = \mathbf{0} \in \mathbb{R}^{d+1}$ and $t^{(0)} = 1$.

- If, at the s-th iteration $(\boldsymbol{\gamma}^{(s)}, t^{(s)})$ satisfies the line-search condition (19), then we update the estimates

$$\begin{aligned}\boldsymbol{\gamma}^{(s+1)} &\leftarrow \text{prox}_{t^{(s)}}(\boldsymbol{\gamma}^{(s)} - t^{(s)}\nabla L_N(\boldsymbol{\gamma}^{(s)})) \\ &= \left(\boldsymbol{\gamma}^{(s)} - t^{(s)}\nabla L_N(\boldsymbol{\gamma}^{(s)})\right)\left(\mathbf{1} - \frac{t^{(s)}\lambda}{\left|\boldsymbol{\gamma}^{(s)} - t^{(s)}\nabla L_N(\boldsymbol{\gamma}^{(s)})\right|}\right)_+,\end{aligned} \tag{21}$$

  and keep the step size unchanged, that is, $t^{(s+1)} \leftarrow t^{(s)}$.

Algorithm 1

Fix a value of $\tau \in (0,1)$ and $\delta > 0$. Initialize with $\boldsymbol{\gamma}^{(0)} = \mathbf{0} \in \mathbb{R}^{d+1}$ and $t^{(0)} = 1$.
while $\|\boldsymbol{\gamma}^{(s+1)} - \boldsymbol{\gamma}^{(s)}\|_1 > \delta$ **do**
if $L_N(\boldsymbol{\gamma}^{(s)} - t^{(s)} G_{t^{(s)}}(\boldsymbol{\gamma}^{(s)})) \leqslant L_N(\boldsymbol{\gamma}^{(s)}) - t^{(s)} \nabla L_N(\boldsymbol{\gamma}^{(s)})^\top G_{t^{(s)}}(\boldsymbol{\gamma}^{(s)}) + \frac{t^{(s)}}{2} \|G_{t^{(s)}}(\boldsymbol{\gamma}^{(s)})\|_2^2$ **then**

$$\boldsymbol{\gamma}^{(s+1)} \leftarrow \mathrm{prox}_{t^{(s)}}(\boldsymbol{\gamma}^{(s)} - t^{(s)} \nabla L_N(\boldsymbol{\gamma}^{(s)})) \text{ and } t^{(s+1)} \leftarrow t^{(s)}.$$

end if
if $L_N(\boldsymbol{\gamma}^{(s)} - t^{(s)} G_{t^{(s)}}(\boldsymbol{\gamma}^{(s)})) > L_N(\boldsymbol{\gamma}^{(s)}) - t^{(s)} \nabla L_N(\boldsymbol{\gamma}^{(s)})^\top G_{t^{(s)}}(\boldsymbol{\gamma}^{(s)}) + \frac{t^{(s)}}{2} \|G_{t^{(s)}}(\boldsymbol{\gamma}^{(s)})\|_2^2$ **then**

$$\boldsymbol{\gamma}^{(s+1)} \leftarrow \boldsymbol{\gamma}^{(s)} \text{ and } t^{(s+1)} \leftarrow \tau t^{(s)}.$$

end if
end while

- If at the s-th iteration $(\boldsymbol{\gamma}^{(s)}, t^{(s)})$ does not satisfy the line-search condition (19) we shrink the step size by a factor of $\tau \in (0,1)$, that is, $t^{(s+1)} \leftarrow \tau t^{(s)}$, and keep the estimates unchanged, that is, $\boldsymbol{\gamma}^{(s+1)} \leftarrow \boldsymbol{\gamma}^{(s)}$.

The procedure is summarized in Algorithm 1.

Note that the smooth part of the objective function L_N is differentiable and its gradient ∇L_N is Lipschitz (by Lemma 24). Hence, Algorithm 1 reaches ε-close to the optimum value in $O(1/\varepsilon)$ iterations (Vandenberghe, 2022).

4.2 Numerical Experiments

We evaluate the performance of the PMPL estimator using Algorithm 1 on synthetic data. The first step is to develop an algorithm to sample from model (3). As mentioned before, direct sampling from the model (3) is computationally challenging due to the presence of an intractable normalizing constant. To circumvent this issue, we deploy a Gibbs sampling algorithm which iteratively updates each outcome variable X_i, for $1 \leqslant i \leqslant N$, based on the conditional distribution $\mathbb{P}(X_i | (X_j)_{j \neq i}, \boldsymbol{Z})$ (recall (4)). Formally, the sampling algorithm can be described as follows:

- Start with an initial configuration $\boldsymbol{X}^{(0)} := (X_i^{(0)})_{1 \leqslant i \leqslant N} \in \{-1, +1\}^N$.
- At the $(s+1)$-th step, for $s \geqslant 1$, choose a vertex of G_N uniformly at random. If the vertex $1 \leqslant i \leqslant N$ is selected, then update $X_i^{(s)}$ to

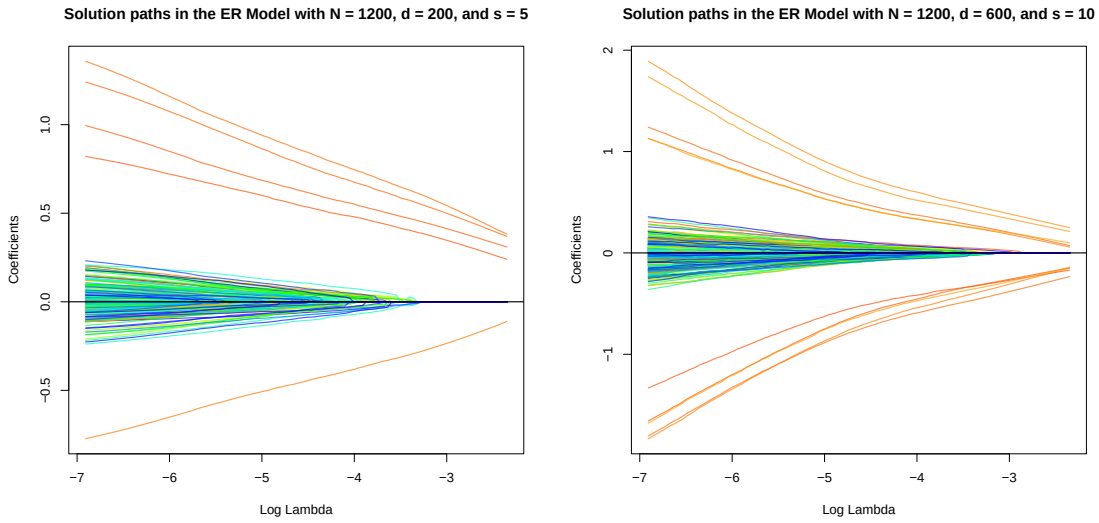
$$X_i^{(s+1)} = \begin{cases} +1, & \text{with probability } \mathbb{P}(X_i^{(s)} = 1 | (X_j^{(s)})_{j \neq i}, \boldsymbol{Z}) \\ -1, & \text{with probability } \mathbb{P}(X_i^{(s)} = -1 | (X_j^{(s)})_{j \neq i}, \boldsymbol{Z}) \end{cases},$$

  and keep $X_j^{(s+1)} = X_j^{(s)}$, for $j \neq i$. Define $\boldsymbol{X}^{(s+1)} := (X_i^{(s+1)})_{1 \leqslant i \leqslant N}$.

The Markov chain $\{\boldsymbol{X}^{(s+1)}\}_{s \geqslant 0}$ has stationary distribution (3) and, hence, can be used to generate approximate samples from (3).

For the numerical experiments, we consider $\beta = 0.3$ and choose the first s regression coefficients $\theta_1, \theta_2, \ldots, \theta_s$ independently from Uniform($[-1, -\frac{1}{2}] \cup [\frac{1}{2}, 1]$), while the remaining $d - s$ regression coefficients $\theta_{s+1}, \theta_{s+2}, \ldots, \theta_d$ are set to zero. The covariates $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N$ are sampled i.i.d. from a d-dimensional multivariate Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = ((\sigma_{ij}))_{1 \leqslant i,j \leqslant d}$, with $\sigma_{ij} = 0.2^{|i-j|}$. With the aforementioned choices of the parameters and the covariates, we generate a sample from the model (3) by running the Gibbs sampling algorithm described above for 30000 iterations. We then apply Algorithm 1 by setting $\varepsilon = 0.001$, $\tau = 0.8$, and consider the solution paths of the PMPL estimator as a function of $\log(\lambda)$, when the network G_N is selected to be the Erdős-Rényi (ER) model and the stochastic block model (SBM). We set the range of λ to be a geometric sequence of length 100 from 0.001 to 0.1.

- Figure 1 depicts the solution paths of the PMPL estimate when G_N is a realization of the Erdős-Rényi random graph $G(N, 5/N)$. Figure 1 (a), corresponds to a setting $N = 1200$, $d = 200$, $s = 5$, while Figure 1 (b) to $N = 1200$, $d = 600$, $s = 600$.

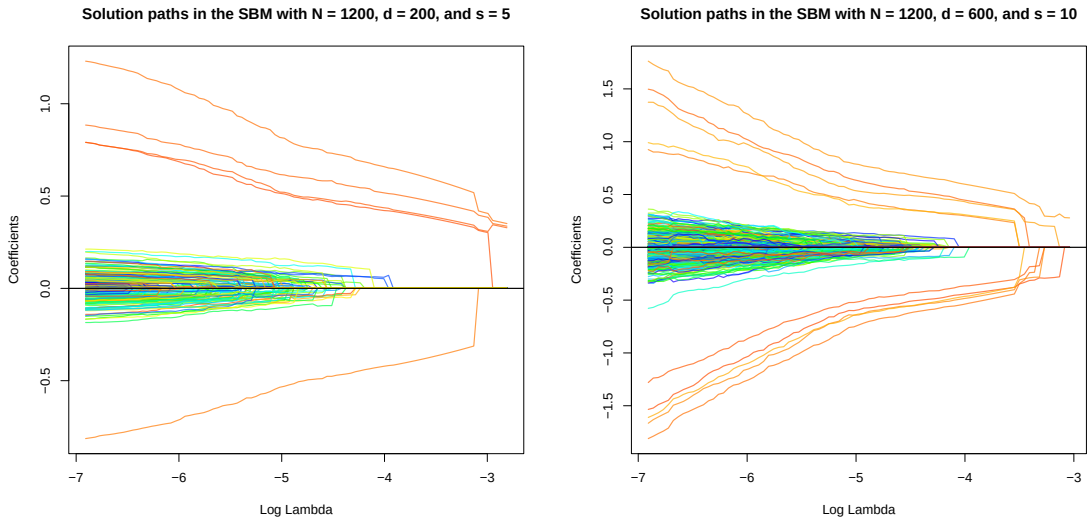


(a) (b)

Figure 1: Solution paths of the PMPL estimates in the Erdős-Rényi model $G(N, 5/N)$: (a) $N = 1200$, $d = 200$, $s = 5$, and (b) $N = 1200$, $d = 600$, $s = 10$.

- Figure 2 shows the solution paths of the PMPL estimate when G_N is a realization of a SBM with $K = 2$, $\lambda_1 = \lambda_2 = \frac{1}{2}$, $p_{11} = p_{22} = 4/N$, and $p_{12} = 8/N$ (that is, a SBM with 2 equal size blocks with within block connection probability $4/N$ and between block connection probability $8/N$ (recall Example 1)). In Figure 2 (a) we have $N = 1200$, $d = 200$, $s = 5$, and in Figure 2 (b) $N = 1200$, $d = 600$, $s = 10$.

From the plots in Figures 1 and 2, it is evident that the first 5 signal (non-zero) coefficients remain non-zero throughout the range of tuning parameters λ considered. Moreover, as



(a) (b)

Figure 2: Solution paths of the PMPL estimates in the stochastic block model: (a) $N = 1200$, $d = 200$, $s = 5$, and (b) $N = 1200$, $d = 600$, $s = 10$.

expected, λ needs to be larger when $d = 600$ for the non-signal (zero) coefficients to shrink to zero exactly.

Next, we investigate the estimation errors by varying the size N of the network G_N. To select the regularization parameter λ, we use a Bayesian Information Criterion (BIC). Specifically, we define

$$\mathrm{BIC}(\lambda) = L_N(\hat{\beta}_\lambda, \hat{\boldsymbol{\theta}}_\lambda) + \mathrm{df}(\lambda) \log N,$$

where $\hat{\beta}_\lambda$, $\hat{\boldsymbol{\theta}}_\lambda = (\hat{\theta}_{\lambda,i})_{1 \leqslant i \leqslant d}$ are the PMPL estimates obtained from (6) for a fixed value of λ and $\mathrm{df}(\lambda) = |\{1 \leqslant i \leqslant d : \hat{\theta}_{\lambda,i} \neq 0\}|$. We choose $\hat{\lambda}$ by minimizing $\mathrm{BIC}(\lambda)$ over a grid of values of λ and denote the corresponding PMPL estimates by $\tilde{\boldsymbol{\gamma}} = (\hat{\beta}_{\hat{\lambda}}, \hat{\boldsymbol{\theta}}_{\hat{\lambda}}^\top)$. Figure 3 shows the average ℓ_1 and ℓ_2 estimation errors $\|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1$ and $\|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2$ and their 1-standard deviation error bars (over 200 repititions) for the Erdős-Rényi (ER) model and the SBM. We refer to these by `IsingL1` and `IsingL2` in Figure 3, respectively. For comparison purposes, we also show the ℓ_1 and ℓ_2 estimation errors for the classical penalized logistic regression (with no interaction term, that is, $\beta = 0$), denoted by `LogisticL1` and `LogisticL2` in Figure 3, respectively. The parameters in the numerical experiment are set as follows: $\beta = 0.3$, $d = 50$, the first $s = 5$ regression coefficients $\theta_1, \theta_2, \ldots, \theta_5$ are independent samples from Uniform$([-1, -\frac{1}{2}] \cup [\frac{1}{2}, 1])$ and the remaining 45 regression coefficients $\theta_6, \theta_7, \ldots, \theta_{50}$ are set to zero. As before, the covariates $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N$ are sampled i.i.d. from a 50-dimensional multivariate Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = ((\sigma_{ij}))_{1 \leqslant i,j \leqslant 100}$, with $\sigma_{ij} = 0.2^{|i-j|}$.

- Figure 3 (a) shows the estimation errors when G_N is a realization of the Erdős-Rényi random graph $G(N, 1/N)$, as N varies from 200 to 1200 over a grid of 6 values.

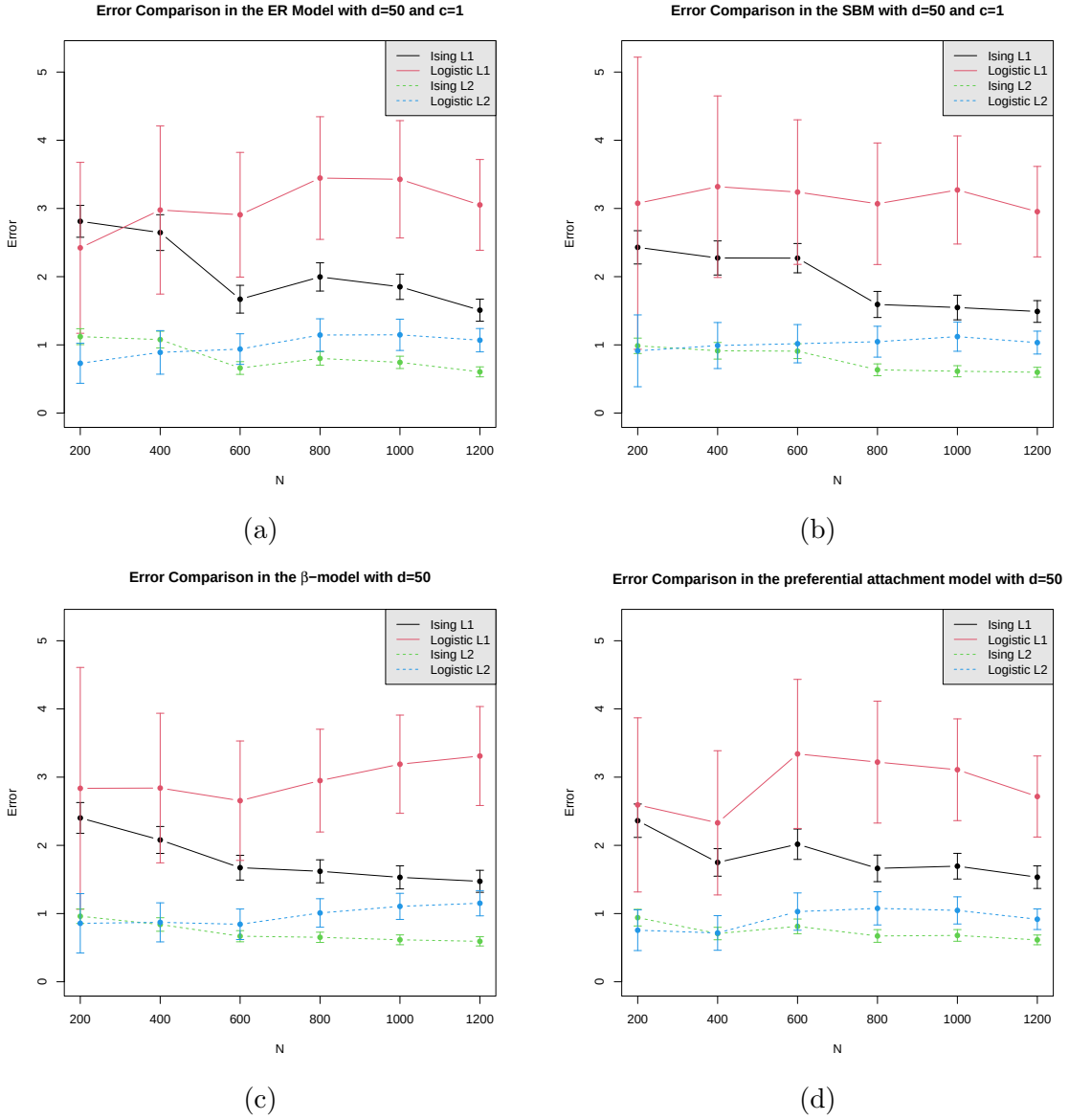


Figure 3: Estimation errors of the PMPL and the penalized logistic regression estimates in the (a) Erdős-Rényi model and (b) the stochastic block model, (c) the β-model, and (d) the preferential attachment model.

- Figure 3 (b) shows the estimation errors when G_N is a realization of a SBM with $K = 2$, $\lambda_1 = \lambda_2 = \frac{1}{2}$, $p_{11} = p_{22} = 0.5/N$, and $p_{12} = 1/N$, with N varying as before.

- Figure 3 (c) shows the estimation errors when G_N is a realization from a β-model (Chatterjee et al., 2011). The β-model is an inhomogeneous random graph model

where each edge (i, j), for $1 \leqslant u < v \leqslant N$, is present independently with probability

$$p_{uv} = \frac{e^{\beta_u+\beta_v}}{1+e^{\beta_u+\beta_v}},$$

with $(\beta_1, \ldots, \beta_N) \in \mathbb{R}^n$. In Figure 3 (c) we chose $\beta_u = -c \cdot u \log(\log(u+1))$, for $1 \leqslant u \leqslant N$, where $c = 200/N$ and N varies from 200 to 1200 over a grid of 6 values.

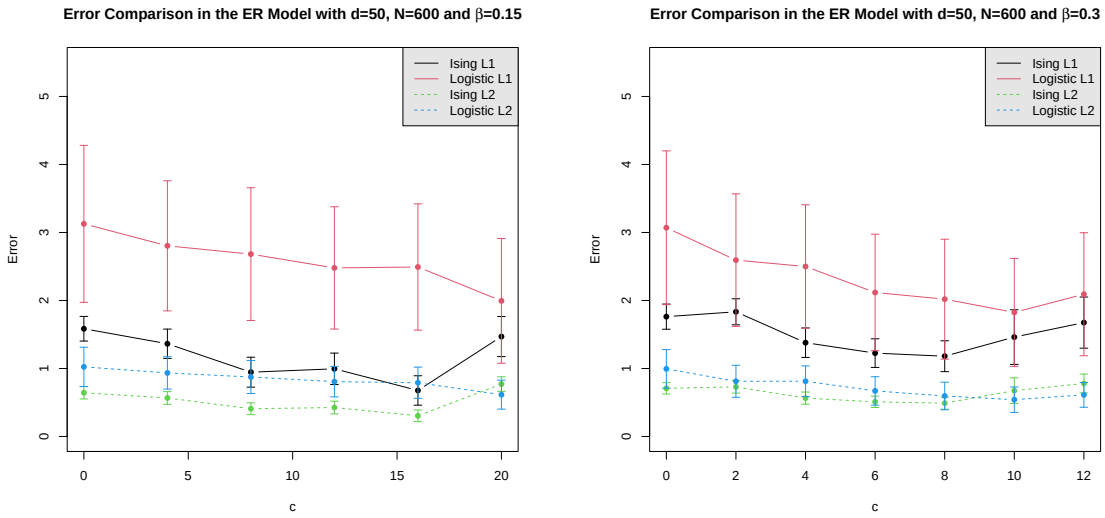
- Figure 3 (d) shows the estimation errors when G_N is a realization from the linear preferential attachment model with one edge added each time, with N varying as before. The linear preferential attachment graph evolves sequentially one vertex at a time, where each new vertex connects to an existing vertex with probability proportional to their degrees (see Bollobas et al. (2003); Krapivsky and Redner (2001)). Consequently, the model exhibits the 'rich gets richer' phenomenon and the degree sequence follows a power law distribution (Barabási and Albert, 1999).

The plots in Figure 3 show that the estimation errors of PMPL estimates exhibit a decreasing trend as N increases, validating the consistency results established in Section 2. Although the ℓ_2 errors of the PMPL and penalized logistic regression estimates are similar for small N, the PMPL errors are better as N increases. Also, the difference between the ℓ_1 errors of the PMPL and penalized logistic regression estimates is significant. While the ℓ_1 errors of PMPL estimates show consistent decreasing trends in all four settings, those for the penalized logistic regression estimates are much higher. Moreover, as expected, the empirical variances of the ℓ_1 and ℓ_2 errors for the penalized logistic regression estimate are significantly larger than those for the PMPL estimate. These findings illustrate the effectiveness of the proposed method for modeling dependent network data for range of network models, encompassing different network topologies, such as community structure and degree distribution.

We also investigate how the PMPL estimate performs with respect to the density of the network. To this end, we consider the Erdős-Rényi random graph $G(N, c/N)$, with $N = 600$, and vary c. Figure 4 shows the estimation errors as c increases, with dependence parameter (a) $\beta = 0.15$ and (b) $\beta = 0.3$ in the respective sub-plots. As expected, the error curves for the PMPL estimates are generally better than those for the penalized logistic regression estimates. Moreover, the estimation errors are relatively small to begin with (when c is small), but starts to show an increasing trend with c after a while. This is expected because as the network density increases the rate of convergence slows down and, as a result, consistent estimation becomes harder (recall the discussion in Section 3.3).

5. Application to Spatial Transcriptomics

In this section, we illustrate how the proposed model can be useful in selecting relevant genes in spatial gene expression data. As mentioned in the Introduction, spatial transcriptomics is a new direction in molecular biology where, in addition to measuring the gene expression levels of individual cells, one also has information about the spatial location of the cells (Eng et al., 2019; Goltsev et al., 2018; Palla et al., 2022; Perkel, 2019). To understand how the spatial location of a cell affects its phenotype, it is natural to consider a model as in (3) with a nearest neighbor graph of the cell locations as the underlying network.



(a) (b)

Figure 4: Estimation errors of the PMPL and the penalized logistic regression estimates in the Erdős-Rényi model $G(N, c/N)$, with $N = 600$, as c varies, where (a) $\beta = 0.15$ and (b) $\beta = 0.3$.

We consider the Visium spatial gene expression data set for human breast cancer (see `https://www.10xgenomics.com/spatial-transcriptomics` for details about the spatial capture technology) available in the Python package `scanpy`. The data set is available at `https://support.10xgenomics.com/spatial-gene-expression/datasets` and can be loaded using the Python command:

```
scanpy.datasets.visium_sge(sample_id='V1_Breast_Cancer_Block_A_Section_1')
```

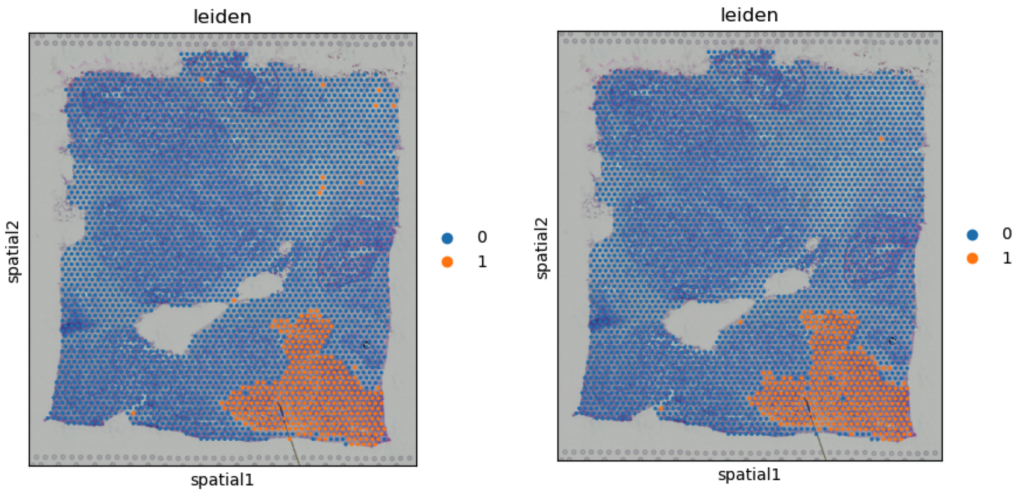
The data consists of 36601 genes and 3798 cells along with their spatial locations. To obtain the cell labels, we first filter out the top 50 highly variable genes, that is, the genes whose expression variance is within the top 50 among all genes. Subsequently, we cluster the cells based on the expression levels of these 50 gene into 2 types (clusters) using the Leiden algorithm (Tragg, 2019). The output of the clustering algorithm visualized using the Python command `sc.pl.spatial` is shown in Figure 5 (a). Using the cell labels obtained as above and the first 100 highly variable genes as the covariates, we then fit the model (3) with the 1-nearest neighbor graph of the spatial location of the cells as the underlying network, using the PMPL method. The optimal λ is chosen using the BIC criterion.

The PMPL method with the BIC chosen regularization parameter, selects 6 genes among the top 100 highly variable genes. Among the selected ones, four of them are actually in the top 50 highly variable gene set obtained in the first filtering step. These genes are shown in Table 1. Next, we re-cluster the cells based on only the 6 selected genes (see Figure 5 (b)). Interestingly, just using the 6 selected genes we can recover the clustering result obtained with the top 50 variable genes with high accuracy. This illustrates how incorporating spatial

information can significantly reduce dimensionality for clustering single cell data and the usefulness of our method in selecting relevant genes.

Selected genes among top 100	Estimated coefficients	Selected genes among top 50
S100A9	0.0087	S100A9
CPB1	-0.0330	CPB1
SPP1	0.0123	SPP1
CRISP3	0.1465	CRISP3
SLITRK6	0.1276	
IGLC2	-0.0392	

Table 1: Names of the selected genes and the estimates of the corresponding regression coefficients. The estimate of β is $\hat{\beta} = 0.1203$.



(a) (b)

Figure 5: Clustering results using the Leiden algorithm: (a) with the top 50 highly variable genes, (b) with the 6 selected genes.

To capture the spatial dependence one can, more generally, consider the K-nearest neighbor graph (instead of the 1-nearest neighbor graph as above) of the spatial locations of the cells in the model (3). To understand the sensitivity of the PMPL method on the choice of the number of nearest neighbors, we repeat the experiment with $K = 1$, $K = 2$, and $K = 3$. The genes selected by the PMPL method and the estimates of the corresponding regression coefficients for each of these settings are shown in Table 2. It turns out that for $K = 1$ and $K = 2$ the genes selected are the same, and for $K = 3$ the genes selected match except one (the gene SPP1 is no longer selected). This shows that the PMPL method is quite robust to choice of the underlying nearest-neighbor graph as long as K is not too large. While one can incorporate more distant spatial dependencies by increasing K, this makes the graph denser and, as a result, the rate of estimation becomes slower (as shown

in Section 3.3). In practice, especially for spatial problems, where the dependence often decreases with distance, choosing a small value of K should suffice.

	Selected genes among top 100	Estimated coefficients	Selected genes among top 50
$K = 1$	S100A9	0.0087	S100A9
	CPB1	-0.0330	CPB1
	SPP1	0.0123	SPP1
	CRISP3	0.1465	CRISP3
	SLITRK6	0.1276	
	IGLC2	-0.0392	
$K = 2$	S100A9	0.0047	S100A9
	CPB1	-0.0250	CPB1
	SPP1	0.0037	SPP1
	CRISP3	0.1121	CRISP3
	SLITRK6	0.1131	
	IGLC2	-0.0317	
$K = 3$	S100A9	0.0034	S100A9
	CPB1	-0.0177	CPB1
	CRISP3	0.0805	CRISP3
	SLITRK6	0.0845	
	IGLC2	-0.0241	

Table 2: Names of the selected genes and the estimates of the corresponding regression coefficients for the K-nearest-neighbor graph, with $K = 1$, $K = 2$, and $K = 3$. The estimates of β are $\hat{\beta} = 0.1203$, $\hat{\beta} = 0.2434$, and $\hat{\beta} = 0.2870$ for $K = 1$, $K = 2$, and $K = 3$, respectively.

6. Conclusion

Understanding the effect of dependence in high-dimensional inference tasks for non-Gaussian models is an emerging research direction. In this paper, we develop a framework for efficient parameter estimation in a model for dependent network data with binary outcomes and high-dimensional covariates. The model combines the classical high-dimensional logistic regression with the Ising model from statistical physics to simultaneously capture dependence from the underlying network and the effect of high-dimensional covariates. This dependence makes the model different and the analysis more challenging compared to existing results based on independent samples. In the this paper we develop an efficient algorithm for jointly estimating the effect of dependence and the high-dimensional regressions parameters using a penalized maximum pseudo-likelihood (PMPL) method and derive its rate of consistency. To understand which of the covariates have an effect on the outcome under the presence of network dependence, we also consider the problem of estimation given a fixed (known) level of dependence. Towards this, we show that using the PMPL method the regression parameters can be estimated at the classical high-dimensional rate, despite the presence of dependence, in the entire high-dimensional regime. We expect the model to be broadly useful in network econometrics and spatial statistics for understanding dependent binary

data with an underlying network geometry. As an application, we apply the proposed model to select genes in spatial transcriptomics data.

Various questions remain and future directions emerge. Theoretically, it would be interesting to see if the conditions for joint estimation can be relaxed. Computationally, it would be interesting to explore more efficient sampling schemes for Ising models with covariates. Incorporating dependence in other generalized linear models and high-dimensional distributions, through the lens of the Ising and more general graphical models, is another interesting direction for future research.

Acknowledgments

Bhaswar B. Bhattacharya thanks Rajarshi Mukherjee and Nancy R. Zhang for many helpful discussions. The authors also thank the anonymous referees for their insightful comments which improved the quality and the presentation of the paper. The authors are also grateful to Sagnik Nandy for his help with the data analysis and Baichen Yu for pointing out an important typo in a previous version of the manuscript. Bhaswar B. Bhattacharya was supported by NSF CAREER grant DMS 2046393 and a Sloan Research Fellowship. Somabha Mukherjee was supported by the National University of Singapore start-up grant WBS A0008523-00-00 and the FoS Tier 1 grant WBS A-8001449-00-00. George Michailidis was supported by NSF grant DMS 2334735.

References

Radosław Adamczak, Michał Kotowski, Bartłomiej Polaczyk, and Michał Strzelecki. A note on concentration for polynomials in the Ising model. *Electronic Journal of Probability*, 24: 1–22, January 2019.

Animashree Anandkumar, Vincent Y. F. Tan, Furong Huang, and Alan S. Willsky. High-dimensional structure estimation in Ising models: Local separation criterion. *The Annals of Statistics*, 40 (3): 1346–1375, June 2012.

Chen Avin, Kaushik Mondal, and Stefan Schmid. Demand-aware network designs of bounded degree. *Distributed Computing*, 33 (3-4): 311–325, June 2020.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, January 2010.

Sudipto Banerjee, Alan E. Gelfand, and Bradley P. Carlin. *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC Press, Boca Raton, Florida, second edition, 2015.

Dhruv Batra, A. C. Gallagher, Devi Parikh, and Tsuhan Chen. Beyond trees: MRF inference via outer-planar decomposition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2496–2503, June 2010.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

Florent Benaych-Georges, Charles Bordenave, and Antti Knowles. Largest eigenvalues of sparse inhomogeneous Erdős-Rényi graphs. *The Annals of Probability*, 47 (3): 1653 - 1676, May 2019.

Marianne Bertrand, Erzo F. P. Luttmer, and Sendhil Mullainathan. Network effects and welfare cultures. *Quarterly Journal of Economics*, 115 (3): 1019–1055, August 2000.

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

Julian Besag. Statistical analysis of non-lattice data. *The Statistician*, 24 (3): 179–195, September 1975.

Bhaswar B. Bhattacharya and Sumit Mukherjee. Inference in Ising models. *Bernoulli*, 24 (1): 493–525, February 2018.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37 (4): 1705–1732, August 2009.

Bela Bollobas, Christian Borgs, Jennifer Chayes, and Oliver Riordan. Directed scale-free graphs. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 132–139, 2003.

Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures and Algorithms*, 31 (1): 3–122, August 2007.

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, Portland Oregon USA, June 2015.

Guy Bresler and Mina Karzand. Learning a tree-structured Ising model in order to make predictions. *The Annals of Statistics*, 48 (2): 713–737, April 2020.

Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+\ell_2$ penalization. *Electronic Journal of Statistics*, 2: 1153–1194, January 2008.

Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35 (6): 2313–2351, December 2007.

Emmanuel J. Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48 (1): 27–42, February 2020.

Yuan Cao, Matey Neykov, and Han Liu. High-temperature structure detection in ferromagnets. *Information and Inference: A Journal of the IMA*, 11 (1): 55–102, March 2022.

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

Sourav Chatterjee. Concentration inequalities with exchangeable pairs (Ph.D. thesis), March 2016. arXiv:math/0507526.

Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. *The Annals of Applied Probability*, 21 (4): 1400–1435, August 2011.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14 (3): 462–467, May 1968.

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

Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6 (2): 125–145, November 2002.

Francis Comets. On consistency of a class of estimators for exponential families of Markov Random Fields on the lattice. *The Annals of Statistics*, 20 (1): 557–578, March 1992.

Francis Comets and Basilis Gidas. Asymptotics of Maximum Likelihood Estimators for the Curie-Weiss Model. *The Annals of Statistics*, 19 (2): 557–578, June 1991.

Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Siddhartha Jayanti. Learning from Weakly Dependent Data under Dobrushin's Condition. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 914–928. PMLR, June 2019.

Yuval Dagan, Constantinos Daskalakis, Nishanth Dikkala, and Anthimos Vardis Kandiros. Learning Ising models from one or multiple samples. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 161–168, June 2021.

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising Models. *IEEE Transactions on Information Theory*, 65 (11): 6829–6852, November 2019.

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

Constantinos Daskalakis, Nishanth Dikkala, and Ioannis Panageas. Logistic regression with peer-group effects via inference in higher-order Ising models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 3653–3663. PMLR, June 2020.

Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer, 2008.

Nabarun Deb, Rajarshi Mukherjee, Sumit Mukherjee, and Ming Yuan. Detecting structured signals in Ising models. *The Annals of Applied Probability*, 34 (1A), 1–45, 2024.

E. Duflo and E. 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, August 2003.

Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, and Long Cai. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568 (7751): 235–239, April 2019.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1), 2010.

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, 1986.

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

B. Gidas. Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions. In Wendell Fleming and Pierre-Louis Lions, editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, The IMA Volumes in Mathematics and Its Applications, pages 129–145, New York, NY, 1988. Springer.

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

Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*, 174 (4): 968–981.e15, August 2018.

Peter J Green and Sylvia Richardson. Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97 (460): 1055–1070, December 2002.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98 (1): 1–15, March 2011.

Xavier Guyon and Hans R. Künsch. Asymptotic comparison of estimators in the Ising model. In Piero Barone, Arnoldo Frigessi, and Mauro Piccioni, editors, *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, Lecture Notes in Statistics, pages 177–198, New York, NY, 1992. Springer.

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*, volume 30: 2463–2472, Inc., 2017.

K. M. Harris, National Longitudinal Study of Adolescent Health, et al. Waves i & ii, 1994-1996; wave iii, 2001-2002; wave iv, 2007-2009.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79 (8): 2554–2558, April 1982.

David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Number 398 in Wiley series in probability and statistics. Wiley, Hoboken, New Jersey, third edition edition, 2013.

Danyang Huang, Wei Lan, Hao Helen Zhang, and Hansheng Wang. Least squares estimation of spatial autoregressive models for large-scale social networks. *Electronic Journal of Statistics*, 13: 1135–1165, 2019.

Ernst Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31 (1): 253–258, February 1925.

Jason K. Johnson, Diane Oyen, Michael Chertkov, and Praneeth Netrapalli. Learning planar ising models. *The Journal of Machine Learning Research*, 17 (1): 7539–7564, January 2016.

Sham Kakade, Ohad Shamir, Karthik Sindharan, and Ambuj Tewari. Learning exponential families in high-dimensions: strong convexity and sparsity. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 381–388. JMLR Workshop and Conference Proceedings, March 2010.

Vardis Kandiros, Yuval Dagan, Nishanth Dikkala, Surbhi Goel, and Constantinos Daskalakis. Statistical estimation from dependent data. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5269–5278. PMLR, July 2021.

Minwoo Kim, Shrijita Bhattacharya, and Tapabrata Maiti. Variational Bayes algorithm and posterior consistency of Ising model parameter estimation, arXiv:2109.01548, September 2021.

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, Berkeley, CA, October 2017.

Paul L Krapivsky and Sidney Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.

Michael Krivelevich and Benny Sudakov. The largest eigenvalue of sparse random graphs. *Combinatorics, Probability and Computing*, 12 (1): 61–72, January 2003.

Erich Leo Lehmann and Joseph P Romano. *Testing statistical hypotheses*, volume 3. Springer, 2005.

Lung-Fei Lee. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):1899–1925, 2004.

Lung-fei Lee, Xiaodong Liu, and Xu Lin. Specification and estimation of social interaction models with network structures. *The Econometrics Journal*, 13(2):145–176, 2010.

Jing Lei. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44 (1): 401–424, February 2016.

Fan Li and Nancy R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105 (491): 1202–1214, September 2010.

Fan Li, Tingting Zhang, Quanli Wang, Marlen Z. Gonzalez, Erin L. Maresh, and James A. Coan. Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *The Annals of Applied Statistics*, 9 (2): 687–713, June 2015.

Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45 (2): 866-896, April 2017.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Springer US, Boston, MA, 1989.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression: Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70 (1): 53–71, January 2008.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37 (1): 246–270, February 2009.

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

Rajarshi Mukherjee and Gourab Ray. On testing for parameters in Ising models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 58 (1): 164–187, February 2022.

Rajarshi Mukherjee, Sumit Mukherjee, and Ming Yuan. Global testing against sparse alternatives under Ising models. *The Annals of Statistics*, 46 (5): 2062–2093, October 2018.

Somabha Mukherjee, Jaesung Son, and Bhaswar B Bhattacharya. Estimation in tensor Ising models. *Information and Inference: A Journal of the IMA*, June 2022.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27 (4): 538–557, November 2012.

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135 (3): 370, 1972.

Matey Neykov and Han Liu. Property testing in high-dimensional Ising models. *The Annals of Statistics*, 47 (5): 2472–2503, October 2019.

Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19: 171–178, January 2022.

Jeffrey M. Perkel. Starfish enterprise: finding RNA patterns in single cells. *Nature*, 572 (7770): 549–551, August 2019.

David K. Pickard. Inference for discrete Markov Fields: The simplest nontrivial case. *Journal of the American Statistical Association*, 82 (397): 90–96, March 1987.

Sean Plummer, Debdeep Pati, and Anirban Bhattacharya. Dynamics of coordinate ascent variational inference: A case study in 2D Ising models. *Entropy*, 22 (11): 1263, November 2020.

Zhiwei Qin, Katya Scheinberg, and Donald Goldfarb. Efficient block-coordinate descent algorithms for the Group Lasso. *Mathematical Programming Computation*, 5 (2): 143–169, June 2013.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57 (10): 6976–6994, October 2011.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38 (3): 1287–1319, June 2010.

B. Sacerdote. Peer effects with random assignment: results for Dartmouth roommates. *The Quarterly Journal of Economics*, 116 (2): 681–704, May 2001.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression, arXiv:1906.03761, November 2019.

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, July 2012.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116 (29): 14516–14525, July 2019.

Pragya Sur, Yuxin Chen, and Emmanuel J. Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled Chi-square. *Probability Theory and Related Fields*, 175(1-2): 487–558, October 2019.

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, and Carey E. Priebe. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23 (3): 1599–1630, August 2017.

V. Traag, L. Waltman, and N. van Eck, From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*. 9, 5233, 2019.

Justin G. Trogdon, James Nonnemaker, and Joanne Pais. Peer effects in adolescent overweight. *Journal of Health Economics*, 27 (5): 1388–1399, September 2008.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998. ISBN 9780521784504.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3): 1166–1202, June 2014.

Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2): 614–645, April 2008.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3: 1166–1202, January 2009.

L. Vandenberghe, Proximal gradient method, ECE236C Lecture Notes, UCLA, 2022. (`http://www.seas.ucla.edu/~vandenbe/236C/lectures/proxgrad.pdf`)

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*, volume 29: 2595–2603, 2016.

Martin Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Number 48 in Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge; New York, NY, 2019.

Bo-Ying Wang and Bo-Yan Xi. Some inequalities for singular values of matrix products. *Linear Algebra and its Applications*, 264: 109–115, October 1997.

Lingzhou Xue, Hui Zou, and Tianxi Cai. Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *The Annals of Statistics*, 40 (3): 1403–1429, June 2012.

Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

Xuening Zhu, Danyang Huang, Rui Pan, and Hansheng Wang. Multivariate spatial autoregressive model for large scale social networks. *Journal of Econometrics*, 215(2):591–606, 2020.

Appendix

Appendix A. Proof of Theorem 1

Theorem 1 is a consequence of the following more general result which provides rates of consistency for the PMPL estimate in terms of $\|\boldsymbol{A}\|_F^2$.

Proposition 7 *Suppose that Assumptions 1, 2, and 3 hold. Then, there exists a constant $\delta > 0$ such that by choosing $\lambda := \delta\sqrt{\log(d+1)/N}$ in the objective function in (22) we have,*

$$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2 = O_s\left(\sqrt{\frac{\log d}{\|\boldsymbol{A}\|_F^4/N}}\right) \quad \text{and} \quad \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 = O_s\left(\sqrt{\frac{\log d}{\|\boldsymbol{A}\|_F^4/N}}\right),$$

with probability $1 - o(1)$, as $N \to \infty$ and $d \to \infty$ such that $d = o(\|\boldsymbol{A}\|_F^2)$ and $\log d = o(\|\boldsymbol{A}\|_F^4/N)$.

Note that when $\liminf_{N\to\infty} \frac{1}{N}\|\boldsymbol{A}\|_F^2 > 0$, then rates in Theorem 7 is an immediate consequence of Proposition 7.

The rest of this section is devoted the proof of Proposition 7. To this end, recall from (6) that our PMPL estimator is defined as:

$$(\hat{\beta}, \hat{\boldsymbol{\theta}}^\top) := \underset{(\beta,\boldsymbol{\theta}^\top)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \; L_N(\beta, \boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1 \tag{22}$$

where $\lambda > 0$ is a tuning parameter and

$$L_N(\beta, \boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^N \left[\log\cosh\left(\beta\sum_{j=1}^N A_{ij}X_j + \boldsymbol{\theta}^\top \boldsymbol{Z}_i\right) - X_i\left(\beta\sum_{j=1}^N A_{ij}X_j + \boldsymbol{\theta}^\top \boldsymbol{Z}_i\right)\right],$$

is as defined in (5) (where we have dropped the additive factor of $\log 2$). To begin with, note that since Assumption 1 holds, by scaling the interaction matrix and the covariate vectors by $\|\boldsymbol{A}\|_\infty$ we can assume without loss of generality,

$$\sup_{N\geqslant 1} \|\boldsymbol{A}\|_\infty \leqslant 1. \tag{23}$$

The first step towards the proof of Theorem 1 is to establish the concentration of the pseudo-likelihood gradient vector $\nabla L_N(\hat{\boldsymbol{\gamma}})$ in the ℓ_∞ norm. This is formalized in the following lemma which is proved in Section A.1.

Lemma 8 (Concentration of the gradient) *For $\hat{\boldsymbol{\gamma}}$ and any $\lambda > 0$,*

$$\|\nabla L_N(\hat{\boldsymbol{\gamma}})\|_\infty \leqslant \lambda. \tag{24}$$

Moreover, there exists $\delta > 0$ such that with $\lambda := \delta\sqrt{\log(d+1)/N}$ the following holds:

$$\mathbb{P}\left(\|\nabla L_N(\boldsymbol{\gamma})\|_\infty > \frac{\lambda}{2}\right) = o(1), \tag{25}$$

where the $o(1)$-term goes to infinity as $d \to \infty$.

The next lemma shows that the pseudo-likelihood function is strongly concave with high probability. The proof of this lemma is given in Section A.2.

Lemma 9 (Strong concavity of pseudo-likelihood) *Suppose the assumptions of Theorem 1 hold. Then, there exists a constant $\kappa := \kappa(s, M, \beta, \Theta) > 0$, such that*

$$L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \geqslant \kappa \frac{\|\boldsymbol{A}\|_F^2 \, \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2}{N},$$

with probability $1 - o(1)$.

The proof of Theorem 1 can now be easily completed using the above lemmas. Towards this define:

$$S := \{1 \leqslant i \leqslant d : \theta_i \neq 0\}.$$

Moreover, for any vector $\boldsymbol{a} \in \mathbb{R}^p$ and any set $Q \subseteq \{1, \ldots, p\}$, we denote by $\boldsymbol{a}_Q$ the vector $(a_i)_{i \in Q}$. Now, for the constant κ as in Lemma 9, consider the event

$$\mathcal{E}_N := \Big\{ \boldsymbol{X} \in \mathcal{C}_N : \|\nabla L_N(\boldsymbol{\gamma})\|_\infty \leqslant \tfrac{\lambda}{2} \\ \text{and } L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \geqslant \kappa \frac{\|\boldsymbol{A}\|_F^2 \, \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2}{N} \Big\}.$$

Clearly, from Lemma 8 and Lemma 9, $\mathbb{P}(\mathcal{E}_N^c) = o(1)$.

Next, suppose $\boldsymbol{X} \in \mathcal{E}_N$. From the definition of $\hat{\boldsymbol{\gamma}}$ it follows that

$$L_N(\hat{\boldsymbol{\gamma}}) + \lambda \|\hat{\boldsymbol{\theta}}\|_1 \leqslant L_N(\boldsymbol{\gamma}) + \lambda \|\boldsymbol{\theta}\|_1. \tag{26}$$

Hence,

$$\begin{aligned} \lambda(\|\boldsymbol{\theta}\|_1 - \|\hat{\boldsymbol{\theta}}\|_1) &\geqslant L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) \\ &= \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + (L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})) \\ &\geqslant -\|\nabla L_N(\boldsymbol{\gamma})\|_\infty \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 + (L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})) \\ &\geqslant -\frac{\lambda \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1}{2} + (L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})), \end{aligned} \tag{27}$$

where the last step uses $\|\nabla L_N(\boldsymbol{\gamma})\|_\infty \leqslant \frac{\lambda}{2}$, for $\boldsymbol{X} \in \mathcal{E}_N$. Next, note that

$$\begin{aligned} \|\hat{\boldsymbol{\theta}}\|_1 = \|\boldsymbol{\theta}_S + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 + \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1 &\geqslant \|\boldsymbol{\theta}_S\|_1 - \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 + \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1 \\ &= \|\boldsymbol{\theta}\|_1 - \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 + \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1. \end{aligned}$$

This implies,

$$\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 - \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1 \geqslant \|\boldsymbol{\theta}\|_1 - \|\hat{\boldsymbol{\theta}}\|_1. \tag{28}$$

Combining (27) and (28) it follows that

$$\begin{aligned} \lambda \left(\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 - \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1 + \frac{\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1}{2} \right) &\geqslant L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ &\geqslant \kappa \frac{\|\boldsymbol{A}\|_F^2 \, \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2}{N}, \end{aligned} \tag{29}$$

where the last inequality uses $L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \geqslant \kappa \frac{\|\boldsymbol{A}\|_F^2 \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2}{N}$, for $\boldsymbol{X} \in \mathcal{E}_N$. Using $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 = \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 + \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1 + |\hat{\beta} - \beta|$ in the LHS of (29) now gives,

$$\begin{aligned}
\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2 &\leqslant \frac{\lambda N}{\kappa \|\boldsymbol{A}\|_F^2} \left(\frac{3\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1}{2} - \frac{\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1}{2} + \frac{|\hat{\beta} - \beta|}{2} \right) \\
&\lesssim_\kappa \frac{\lambda N}{\|\boldsymbol{A}\|_F^2} \left(\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 + |\hat{\beta} - \beta| \right) \\
&\lesssim_\kappa \frac{\lambda N}{\|\boldsymbol{A}\|_F^2} \sqrt{s+1} \sqrt{\sum_{i \in S} (\hat{\theta}_i - \theta_i)^2 + (\hat{\beta} - \beta)^2} \\
&\lesssim \frac{\lambda N}{\|\boldsymbol{A}\|_F^2} \sqrt{s} \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2.
\end{aligned}$$

This implies, for $\boldsymbol{X} \in \mathcal{E}_N$,

$$||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}||_2 = O_{\kappa,\delta} \left(\frac{N}{\|\boldsymbol{A}\|_F^2} \sqrt{\frac{s \log d}{N}} \right).$$

This completes the proof of the ℓ_2 error rate in Theorem 1, since $\mathbb{P}(\mathcal{E}_N) = 1 - o(1)$. The bound on the ℓ_1 error $||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}||_1$ is shown in Lemma 15.

A.1 Proof of Lemma 8

First, we establish that $||\nabla L_N(\hat{\boldsymbol{\gamma}})||_\infty \leqslant \lambda$. Fix $1 \leqslant i \leqslant d$ and define the univariate function:

$$f(x) := L_N(\hat{\beta}, \hat{\theta}_1, \ldots, \hat{\theta}_{i-1}, x, \hat{\theta}_{i+1}, \ldots, \hat{\theta}_d).$$

Note that $f'(\hat{\theta}_i) = \frac{\partial}{\partial \underline{\theta}_i} \nabla L_N(\underline{\gamma}) \mid_{\underline{\gamma} = \hat{\boldsymbol{\gamma}}}$. Now, by the definition of $\hat{\boldsymbol{\gamma}}$ we have, $f(\hat{\theta}_i) + \lambda|\hat{\theta}_i| \leqslant f(x) + \lambda|x|$, which implies,

$$f(x) - f(\hat{\theta}_i) \geqslant \lambda(|\hat{\theta}_i| - |x|).$$

Then consider the following cases:

- $\hat{\theta}_i > 0$: Then, for all $x > \hat{\theta}_i$,

  $$\frac{f(x) - f(\hat{\theta}_i)}{x - \hat{\theta}_i} \geqslant \lambda \frac{|\hat{\theta}_i| - |x|}{x - \hat{\theta}_i} = -\lambda.$$

  Similarly, for all $0 < x < \hat{\theta}_i$, $\frac{f(x) - f(\hat{\theta}_i)}{x - \hat{\theta}_i} \leqslant \lambda \frac{|\hat{\theta}_i| - |x|}{x - \hat{\theta}_i} = -\lambda$. This implies, $f'(\hat{\theta}_i) = -\lambda$.

- $\hat{\theta}_i < 0$: Then, for all $0 > x > \hat{\theta}_i$,

  $$\frac{f(x) - f(\hat{\theta}_i)}{x - \hat{\theta}_i} \geqslant \lambda \frac{|\hat{\theta}_i| - |x|}{x - \hat{\theta}_i} = \lambda.$$

  Similarly, $x < \hat{\theta}_i$, $\frac{f(x) - f(\hat{\theta}_i)}{x - \hat{\theta}_i} \leqslant \lambda \frac{|\hat{\theta}_i| - |x|}{x - \hat{\theta}_i} = \lambda$. Hence, in this case, $f'(\hat{\theta}_i) = \lambda$.

- $\hat{\theta}_i = 0$: In this case, for all $x > 0$,

$$\frac{f(x) - f(0)}{x} \geqslant -\lambda \frac{|x|}{x} = -\lambda$$

and for all $x < 0$, $\frac{f(x)-f(0)}{x} \leqslant -\lambda \frac{|x|}{x} = \lambda$. Since f' exists, this implies that $|f'(0)| \leqslant \lambda$.

Next, define

$$g(x) := L_N(x, \hat{\theta}_1, \ldots, \hat{\theta}_{i-1}, \hat{\theta}_i, \hat{\theta}_{i+1}, \ldots, \hat{\theta}_d).$$

Note that $g'(\hat{\beta}) = \frac{\partial}{\partial \underline{\beta}} \nabla L_N(\underline{\gamma}) \mid_{\underline{\gamma}=\hat{\gamma}}$. By the definition of $\hat{\gamma}$ we have, $g(\hat{\beta}) \leqslant g(x)$. This implies that $g'(\hat{\beta}) = 0$. Combining the above, it follows that

$$||\nabla L_N(\hat{\gamma})||_\infty = \max_{j \in [d]} |f'(\hat{\theta}_i)| \leqslant \lambda,$$

completing the proof of (24).

Next, we establish the concentration of $\|\nabla L_N(\boldsymbol{\gamma})\|_\infty$ as in (25). For this step, we require the following definitions. For $1 \leqslant i \leqslant N$, denote

$$m_i(\boldsymbol{X}) := \sum_{j=1}^{N} a_{ij} X_j.$$

Define functions $\phi_i : \mathcal{C}_N \to \mathbb{R}$, for $1 \leqslant i \leqslant N$, as follows:

$$\phi_i(\boldsymbol{x}) := -\frac{1}{N} \left\{ m_i(\boldsymbol{x}) \left(x_i - \tanh(\beta m_i(\boldsymbol{x}) + \boldsymbol{\theta}^\top \boldsymbol{Z}_i) \right) \right\}, \tag{30}$$

for $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{C}_N$. Similarly, define functions $\phi_{i,s} : \mathcal{C}_N \to \mathbb{R}$, for $1 \leqslant i \leqslant N$ and $1 \leqslant s \leqslant d$, as follows:

$$\phi_{i,s}(\boldsymbol{x}) := -\frac{1}{N} \left\{ Z_{i,s} \left(x_i - \tanh(\beta m_i(\boldsymbol{x}) + \boldsymbol{\theta}^\top \boldsymbol{Z}_i) \right) \right\}. \tag{31}$$

Note that $\nabla L_N = (\frac{\partial L_N}{\partial \beta}, \frac{\partial L_N}{\partial \theta_1}, \ldots, \frac{\partial L_N}{\partial \theta_d})^\top$ where

$$\frac{\partial L_N}{\partial \beta} = \sum_{i=1}^{N} \phi_i(\boldsymbol{X}) \quad \text{and} \quad \frac{\partial L_N}{\partial \theta_s} = \sum_{i=1}^{N} \phi_{i,s}(\boldsymbol{X}), \quad \text{for } 1 \leqslant s \leqslant d.$$

To establish the concentration of $\|\nabla L_N(\boldsymbol{\gamma})\|_\infty$, we use the conditioning trick from Dagan et al. (2021), which allows to reduce the model (3) to an Ising model in the Dobrushin regime (where the correlations are sufficiently weak and the model approximately behaves like a product measure), by conditioning on a subset of the nodes. To describe this, we need the following definition:

Definition 10 *Suppose that $\boldsymbol{\sigma} \in \{-1, 1\}^N$ is a sample from the Ising model:*

$$\mathbb{P}_{\beta, \boldsymbol{h}}(\boldsymbol{\sigma}) \;\propto\; \exp\left(\boldsymbol{\sigma}^\top \boldsymbol{D} \boldsymbol{\sigma} + \sum_{i=1}^{N} h_i \sigma_i \right), \tag{32}$$

where $\boldsymbol{h} = (h_1, h_2, \ldots, h_n)^\top \in \mathbb{R}^n$ *and* $\boldsymbol{D}$ *is a symmetric matrix with zeros on the diagonal with* $\sup_{N \geqslant 1} \|\boldsymbol{D}\|_\infty \leqslant R$. *Moreover, suppose that with probability* 1,

$$\min_{1 \leqslant i \leqslant N} \mathrm{Var}(\sigma_i | \boldsymbol{\sigma}_{-i}) \geqslant \Upsilon,$$

for some $\Upsilon \geqslant 0$. *Then, the model* (32) *is referred to as an* (R, Υ)*-Ising model.*

Recently, Dagan et al. (2021) developed a technique for reducing an (R, Υ)-Ising model to an (η, Υ)-Ising model, for $0 < \eta < R$, by conditioning on a subset of vertices. As a consequence, by choosing η one can ensure that the conditional model is in the Dobrushin high-temperature regime. Although the Ising model studied in Dagan et al. (2021) is different, the same proof extends to our model (3) as well. We formalize this in the following lemma, which is proved in Appendix D.2.

Lemma 11 *Fix* $R > 0$ *and* $\eta \in (0, R)$. *Let* $\boldsymbol{X} \in \{-1, 1\}^N$ *be a sample from an* (R, Υ)*-Ising model. Then there exist subsets* $I_1, \ldots, I_\ell \subseteq [N]$ *with* $\ell \lesssim R^2 \log N / \eta^2$ *such that:*

1. *For all* $1 \leqslant i \leqslant N$,
$$|\{j \in [\ell] : i \in I_j\}| = \lceil \eta \ell / 8R \rceil$$
2. *For all* $1 \leqslant j \leqslant \ell$, *the conditional distribution of* $\boldsymbol{X}_{I_j}$ *given* $\boldsymbol{X}_{I_j^c} := (X_u)_{u \in [N] \setminus I_j}$ *is an* (η, Υ)*- Ising model.*

Furthermore, for any non-negative vector $\boldsymbol{a} \in \mathbb{R}^N$, *there exists* $j \in \ell$ *such that*

$$\sum_{i \in I_j} a_i \geqslant \frac{\eta}{8R} \sum_{i=1}^N a_i. \tag{33}$$

We will apply the above result to our model (3). Towards this, set $\boldsymbol{D} = \frac{\beta}{2}\boldsymbol{A}$ and $h_i = \boldsymbol{\theta}^\top \boldsymbol{Z}_i$, for $1 \leqslant i \leqslant N$ in (32). Under this parametrization, (3) is an (R, Υ)-Ising model as shown below:

Lemma 12 *The model* (3) *is a* (R, Υ)*-Ising model with* $R = |\beta|/2$ *and any* $\Upsilon = e^{-4\Theta M s - 4|\beta|}$.

Proof Note that for every $j \in [N]$,

$$\mathrm{Var}(X_j | \boldsymbol{X}_{[N] \setminus \{j\}}) = 4p(1-p),$$

where $p = \mathbb{P}(X_j = 1 | \boldsymbol{X}_{[N] \setminus \{j\}})$. Now, denoting the elements of of the matrix $\boldsymbol{D}$ as $(d_{ij})_{1 \leqslant i,j \leqslant N}$, note that

$$p = \frac{\exp\left(h_j + 2\sum_{v \in [N] \setminus \{j\}} d_{jv} X_v\right)}{2\cosh\left(h_j + 2\sum_{v \in [N] \setminus \{j\}} d_{jv} X_v\right)}.$$

Then using the inequality $\frac{e^x}{2\cosh(x)} \geqslant \frac{1}{2} e^{-2|x|}$ gives,

$$\min\{p, 1-p\} \geqslant \frac{1}{2} \exp\left(-2 \left| h_j + 2 \sum_{v \in [N] \setminus \{j\}} d_{jv} X_v \right| \right). \tag{34}$$

Next, using

$$\left| h_j + 2 \sum_{v \in [N] \setminus \{j\}} d_{jv} X_v \right| = \left| \boldsymbol{\theta}^\top \boldsymbol{Z}_j + 2 \sum_{v \in [N] \setminus \{j\}} d_{jv} X_v \right| \leqslant |\boldsymbol{\theta}^\top \boldsymbol{Z}_j| + 2\|\boldsymbol{D}\|_\infty \leqslant \Theta M s + |\beta|,$$

it follows from (34) that

$$\min\{p, 1-p\} \geqslant \frac{1}{2} e^{-2\Theta M s - 2|\beta|}.$$

Hence, we have:

$$\mathrm{Var}(X_j | \boldsymbol{X}_{[N] \setminus \{j\}}) \geqslant e^{-4\Theta M s - 4|\beta|}.$$

This completes the proof of the lemma, since $\|\boldsymbol{D}\|_\infty \leqslant |\beta|$ (since $\|\boldsymbol{A}\|_\infty \leqslant 1$ by (23)). ■

By the above lemma, model (3) is an (R, Υ)-Ising model, with $R := |\beta|/2$ and $\Upsilon = e^{-4\Theta M s - 4|\beta|}$. Next, choose

$$\eta := \min\left\{ \frac{1}{16}, \frac{|\beta|}{2} \right\},$$

and suppose $I_1, \ldots, I_\ell$ are subsets of $[N]$ as in Lemma 11. Then, defining $\ell' := \lceil \eta \ell / R \rceil$ we get,

$$\left| \sum_{i=1}^N \phi_i(\boldsymbol{X}) \right| = \left| \frac{1}{\ell'} \sum_{r=1}^{\ell} \sum_{i \in I_r} \phi_i(\boldsymbol{X}) \right| \leqslant \frac{1}{\ell'} \sum_{r=1}^{\ell} \left| \sum_{i \in I_r} \phi_i(\boldsymbol{X}) \right| \leqslant \frac{\ell}{\ell'} \max_{r \in [\ell]} |Q_r(\boldsymbol{X})|, \tag{35}$$

where

$$Q_r(\boldsymbol{X}) := \sum_{i \in I_r} \phi_i(\boldsymbol{X}), \tag{36}$$

for $r \in [\ell]$. Similarly, it follows that, for $s \in [d]$,

$$\left| \sum_{i=1}^N \phi_{i,s}(\boldsymbol{X}) \right| = \left| \frac{1}{\ell'} \sum_{r=1}^{\ell} \sum_{i \in I_r} \phi_{i,s}(\boldsymbol{X}) \right| \leqslant \frac{1}{\ell'} \sum_{r=1}^{\ell} \left| \sum_{i \in I_r} \phi_{i,s}(\boldsymbol{X}) \right| \leqslant \frac{\ell}{\ell'} \max_{r \in [\ell]} |Q_{r,s}(\boldsymbol{X})|, \tag{37}$$

where

$$Q_{r,s}(\boldsymbol{X}) := \sum_{i \in I_r} \phi_{i,s}(\boldsymbol{X}). \tag{38}$$

The following lemma shows that the functions Q_r and $Q_{r,s}$ are Lipschitz in the Hamming metric. The proof is given in Appendix D.1.

Lemma 13 *For $r \in [\ell]$ and $s \in [d]$, let Q_r and $Q_{r,s}$ be as defined in (36) and (38), respectively. Then for any two vectors $\boldsymbol{X}, \boldsymbol{X}' \in \mathcal{C}_N$ differing in just the k-th coordinate, the following hold:*

(1) For $r \in [\ell]$,

$$|Q_r(\boldsymbol{X}) - Q_r(\boldsymbol{X}')| \leqslant \frac{2|\beta| + 6}{N}.$$

(2) Similarly, for $r \in [\ell]$ and $s \in [d]$,

$$|Q_{r,s}(\boldsymbol{X}) - Q_{r,s}(\boldsymbol{X}')| \leqslant \frac{2|Z_{k,s}|}{N} + \frac{2|\beta|}{N}\sum_{i=1}^{N}|Z_{i,s}a_{ik}| =: c_k.$$

Using the above result together with Lemma 12, we can now establish the concentrations of $Q_r(\boldsymbol{X})$ and $Q_{r,s}(\boldsymbol{X})$, conditional on $\boldsymbol{X}_{I_r^c}$. To this end, recalling the definition of $\phi_i(\cdot)$ from (30) note that $\mathbb{E}\left(Q_r(\boldsymbol{X})|\boldsymbol{X}_{I_r^c}\right) = 0$, for $r \in [\ell]$. Moreover, by Lemma 12, $\boldsymbol{X}|\boldsymbol{X}_{I_r^c}$ is an (η, Υ)-Ising model, where $\eta \leqslant \frac{1}{16}$. Therefore, since Q_r is $O_\beta(1/N)$-Lipschitz (by Lemma 13), applying Theorem 4.3 and Lemma 4.4 in Chatterjee (2016) gives, for every $t > 0$,

$$\mathbb{P}\left(|Q_r(\boldsymbol{X})| \geqslant t \,\middle|\, \boldsymbol{X}_{I_r^c}\right) \leqslant 2e^{-O_\beta(Nt^2)}. \tag{39}$$

Similarly, recalling (31), it follows that for each $r \in [\ell]$ and $s \in [d]$, $\mathbb{E}(Q_{r,s}(\boldsymbol{X})|\boldsymbol{X}_{I_r^c}) = 0$. Then, since $\sum_{k=1}^{N} c_k^2 = O_M(1)$ under Assumptions 1 and 4, Lemma 13 together with Theorem 4.3 and Lemma 4.4 in Chatterjee (2016) gives, for every $t \geqslant 0$,

$$\mathbb{P}\left(|Q_{r,s}(\boldsymbol{X})| \geqslant t \,\middle|\, \boldsymbol{X}_{I_r^c}\right) \leqslant 2e^{-O_{\beta,M}(Nt^2)}. \tag{40}$$

Hence, combining (35), (39), (37), (40), and Lemma 11 (which implies that $\ell = O(\log N)$) gives,

$$\mathbb{P}\left(\left|\sum_{i=1}^{N}\phi_i(\boldsymbol{X})\right| \geqslant t\right) \leqslant 2e^{-O_\beta(Nt^2)} \quad \text{and} \quad \mathbb{P}\left(\left|\sum_{i=1}^{N}\phi_{i,s}(\boldsymbol{X})\right| \geqslant t\right) \leqslant 2e^{-O_{\beta,M}(Nt^2)}, \tag{41}$$

for each $s \in [d]$. It thus follows from (41) and a union bound, that

$$\mathbb{P}\left(\|\nabla L_N(\boldsymbol{\gamma})\|_\infty \geqslant t\right) \leqslant 2(d+1)e^{-KNt^2}, \tag{42}$$

for some constant $K > 0$, depending on β and M. Now, choosing $t = \frac{\lambda}{2} = \frac{1}{2}\delta\sqrt{\log(d+1)/N}$ in (42) above gives,

$$\mathbb{P}\left(\|\nabla L_N(\boldsymbol{\gamma})\|_\infty \geqslant \frac{\lambda}{2}\right) \leqslant 2(d+1)^{1-\frac{K\delta^2}{4}} = o(1),$$

whenever $\delta^2 < 4/K$. This completes the proof of Lemma 8.

A.2 Proof of Lemma 9

Define the following $(d+1) \times (d+1)$ dimensional matrix,

$$\boldsymbol{G} := \frac{1}{N}\begin{pmatrix} \boldsymbol{m}^\top\boldsymbol{m} & \boldsymbol{m}^\top\boldsymbol{Z} \\ \boldsymbol{Z}^\top\boldsymbol{m} & \boldsymbol{Z}^\top\boldsymbol{Z} \end{pmatrix} \tag{43}$$

The key step towards proving Lemma 9 is to show that the lowest eigenvalue of $\nabla^2 L_N$ is bounded away from 0 with high probability.

Lemma 14 *There exists a constant $C > 0$ (depending only on s, Θ and M), such that*

$$\mathbb{P}\left(\lambda_{\min}(\boldsymbol{G}) \geqslant \frac{C\|\boldsymbol{A}\|_F^2}{N}\right) \geqslant 1 - e^{-\Omega(\|\boldsymbol{A}\|_F^4/N)}, \tag{44}$$

as $N, d \to \infty$, such that $d = o(\|\boldsymbol{A}\|_F^2)$.

The proof of Lemma 14 is given in Section A.2.1. We first show it can be used to complete the proof of Lemma 9. To this end, by a second order Taylor series expansion, we know that there exists $\alpha \in (0,1)$ and $\underline{\boldsymbol{\gamma}} = (\underline{\beta}, \underline{\boldsymbol{\theta}}^\top)^\top = \alpha\boldsymbol{\gamma} + (1-\alpha)\hat{\boldsymbol{\gamma}}$ such that

$$\begin{aligned} L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) &= \frac{1}{2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \nabla^2 L_N(\underline{\boldsymbol{\gamma}})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ &= \frac{1}{2N}\sum_{i=1}^N \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{U}_i \boldsymbol{U}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\cosh^2(\underline{\beta} m_i(\boldsymbol{X}) + \underline{\boldsymbol{\theta}}^\top \boldsymbol{Z}_i)}, \end{aligned} \tag{45}$$

where $\boldsymbol{U}_i := (m_i(\boldsymbol{X}), \boldsymbol{Z}_i^\top)^\top$. Now, note that:

$$|\underline{\beta} m_i(\boldsymbol{X})| \leqslant |\underline{\beta}||m_i(\boldsymbol{X})| \leqslant |\underline{\beta}|||A||_\infty \leqslant |\beta| + |\hat{\beta} - \beta| \leqslant |\beta| + \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \tag{46}$$

and

$$|\underline{\boldsymbol{\theta}}^\top \boldsymbol{Z}_i| \leqslant \|\underline{\boldsymbol{\theta}}\|_1 \|\boldsymbol{Z}_i\|_\infty \leqslant M\left(\|\underline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 + \|\boldsymbol{\theta}\|_1\right) \leqslant M(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 + s\Theta). \tag{47}$$

Since cosh is an even function and increasing on the positive axis, we obtain

$$\frac{1}{2N}\sum_{i=1}^N \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{U}_i \boldsymbol{U}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\cosh^2(\underline{\beta} m_i(\boldsymbol{X}) + \underline{\boldsymbol{\theta}}^\top \boldsymbol{Z}_i)} \geqslant \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{G}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{2\cosh^2(|\beta| + (M+1)(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1) + sM\Theta)}, \tag{48}$$

where $\boldsymbol{m} := (m_1(\boldsymbol{X}), \ldots, m_N(\boldsymbol{X}))^\top$, $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)^\top$, and $\boldsymbol{G}$ is as defined in (43). Combining (45) and (48) gives,

$$\begin{aligned} L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) &= \frac{1}{2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \nabla^2 L_N(\underline{\boldsymbol{\gamma}})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ &\geqslant \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{G}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{2\cosh^2(|\beta| + (M+1)(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1) + sM\Theta)}. \end{aligned} \tag{49}$$

Next, we establish a high probability upper bound on $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1$ whenever the conditions of Theorem 1 are satisfied.

Lemma 15 *Suppose (44) holds. Then, for $\lambda := \delta\sqrt{\log(d+1)/N}$ as in Lemma 8,*

$$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 = O_s\left(\frac{N}{\|\boldsymbol{A}\|_F^2}\sqrt{\frac{\log(d+1)}{N}}\right).$$

with probability $1 - o(1)$, whenever $N, d \to \infty$ such that $\log d = o(\|\boldsymbol{A}\|_F^4/N)$,

Proof By the convexity of the function L_N it follows from (26) that

$$\begin{aligned}\lambda(\|\boldsymbol{\theta}\|_1 - \|\hat{\boldsymbol{\theta}}\|_1) &\geqslant L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) \\ &\geqslant \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ &\geqslant -\|\nabla L_N(\boldsymbol{\gamma})\|_\infty \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \\ &\geqslant -\frac{\lambda\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1}{2}, \end{aligned} \tag{50}$$

where the last step uses $\|\nabla L_N(\boldsymbol{\gamma})\|_\infty \leqslant \frac{\lambda}{2}$, for $\boldsymbol{X} \in \mathcal{E}_N$. Recall from (28) that $\|\hat{\boldsymbol{\theta}}\|_1 - \|\boldsymbol{\theta}\|_1 \geqslant \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1 - \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1$. Therefore, from (50), we have:

$$\begin{aligned}\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1 - \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 \leqslant \|\hat{\boldsymbol{\theta}}\|_1 - \|\boldsymbol{\theta}\|_1 &\leqslant \frac{\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1}{2} \\ &= \frac{\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1}{2} + \frac{\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1}{2} + \frac{|\hat{\beta} - \beta|}{2}.\end{aligned}$$

This means, $\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_{S^c}\|_1 \leqslant 3(\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 + |\hat{\beta} - \beta|)$, and hence,

$$\|(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_1 \leqslant 4(\|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 + |\hat{\beta} - \beta|). \tag{51}$$

Denote $\mathcal{K} := \|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_S\|_1 + |\hat{\beta} - \beta|$. By the Cauchy-Schwarz inequality,

$$\mathcal{K} \leqslant \sqrt{s+1}\sqrt{\sum_{i \in S}(\hat{\theta}_i - \theta_i)^2 + (\hat{\beta} - \beta)^2} \leqslant \sqrt{s+1}||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}||_2. \tag{52}$$

Next, for $t \in [0,1]$, let $\boldsymbol{\gamma}_t := t\hat{\boldsymbol{\gamma}} + (1-t)\boldsymbol{\gamma}$, and $g(t) := (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \nabla L_N(\boldsymbol{\gamma}_t)$. Then

$$|g(1) - g(0)| = \left|(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top(\nabla L_N(\hat{\boldsymbol{\gamma}}) - \nabla L_N(\boldsymbol{\gamma}))\right| \leqslant \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \cdot \|\nabla L_N(\hat{\boldsymbol{\gamma}}) - \nabla L_N(\boldsymbol{\gamma})\|_\infty. \tag{53}$$

Therefore,

$$\begin{aligned} g'(t) &= (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \nabla^2 L_N(\boldsymbol{\gamma}_t)(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\ &= \frac{1}{N}\sum_{i=1}^N \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{U}_i \boldsymbol{U}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\cosh^2(\beta_t m_i(\boldsymbol{X}) + \boldsymbol{\theta}_t^\top \boldsymbol{Z}_i)} && (\text{where } \boldsymbol{U}_i := (m_i(\boldsymbol{X}), \boldsymbol{Z}_i^\top)^\top) \\ &\geqslant \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{G}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\cosh^2(|\beta| + (M+1)\|\boldsymbol{\gamma}_t - \boldsymbol{\gamma}\|_1 + sM\Theta)} && (\text{by (46) and (47)}) \\ &\geqslant \frac{\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2 \ \lambda_{\min}(\boldsymbol{G})}{\cosh^2(|\beta| + (M+1)\|\boldsymbol{\gamma}_t - \boldsymbol{\gamma}\|_1 + sM\Theta)} \\ &\gtrsim_C \frac{\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2 \|\boldsymbol{A}\|_F^2}{N\cosh^2(|\beta| + 4|t|(M+1)\mathcal{K} + Ms\Theta)}, \end{aligned}$$

where the last step uses (44) (which holds with probability $1-o(1)$), C is as in Lemma 14, and $\|\boldsymbol{\gamma}_t-\boldsymbol{\gamma}\|_1=|t|\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_1\leqslant 4|t|\mathcal{K}$ (by (51)). Hence,

$$\begin{aligned}|g(1)-g(0)|\geqslant g(1)-g(0)&=\int_0^1 g'(t)\ dt\\&\geqslant\int_0^{\min\{1,\ 1/\mathcal{K}\}} g'(t)\ dt\\&\gtrsim_s \frac{\|\boldsymbol{A}\|_F^2}{N}\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_2^2\ \min\{1,\ 1/\mathcal{K}\}.\end{aligned}\tag{54}$$

Combining (53) and (54) gives,

$$\begin{aligned}\min\{\mathcal{K},\ 1\}&\lesssim_s \frac{\mathcal{K}N\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1}{\|\boldsymbol{A}\|_F^2\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_2^2}\cdot\|\nabla L_N(\hat{\boldsymbol{\gamma}})-\nabla L_N(\boldsymbol{\gamma})\|_\infty\\&\lesssim\frac{\mathcal{K}^2N}{\|\boldsymbol{A}\|_F^2\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_2^2}\cdot\|\nabla L_N(\hat{\boldsymbol{\gamma}})-\nabla L_N(\boldsymbol{\gamma})\|_\infty\qquad\text{(by (51))}\\&\lesssim\frac{sN}{\|\boldsymbol{A}\|_F^2}\|\nabla L_N(\hat{\boldsymbol{\gamma}})-\nabla L_N(\boldsymbol{\gamma})\|_\infty,\end{aligned}\tag{55}$$

using (52). Now, recall that, by Lemma 8, with probability $1-o(1)$, $\|\nabla L_N(\boldsymbol{\gamma})\|_\infty\lesssim_\delta\sqrt{\log(d+1)/N}$ and $\|\nabla L_N(\hat{\boldsymbol{\gamma}})\|_\infty\lesssim\sqrt{\log(d+1)/N}$. Applying this in (55) shows that with probability $1-o(1)$,

$$\min\{\mathcal{K},\ 1\}=O_s\left(\frac{N}{\|\boldsymbol{A}\|_F^2}\sqrt{\frac{\log(d+1)}{N}}\right).$$

This implies,

$$\mathcal{K}=\|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})_S\|_1+|\hat{\beta}-\beta|=O_s\left(\frac{N}{\|\boldsymbol{A}\|_F^2}\sqrt{\frac{\log(d+1)}{N}}\right),$$

with probability $1-o(1)$, whenever $N,d\to\infty$ such that $\log d=o(\|\boldsymbol{A}\|_F^4/N)$. Therefore, by (51), $\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1=O_s\left(\frac{N}{\|\boldsymbol{A}\|_F^2}\sqrt{\frac{\log(d+1)}{N}}\right)$ with probability $1-o(1)$. ■

Using Lemma 14 and Lemma 15 in (49) it follows that, there exists κ (as the statement of Lemma 9) such that

$$\begin{aligned}L_N(\hat{\boldsymbol{\gamma}})-L_N(\boldsymbol{\gamma})-\nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})&\geqslant\frac{(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^\top\boldsymbol{G}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})}{2\cosh^2(|\beta|+(M+1)\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1+sM\Theta)}\\&\gtrsim_{\beta,M,\kappa}\frac{\|\boldsymbol{A}\|_F^2\,\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_2^2}{N}\end{aligned}$$

with probability $1-o(1)$, as $N,d\to\infty$ such that $d=o(\|\boldsymbol{A}\|_F^2)$ and $\log d=o(\|\boldsymbol{A}\|_F^4/N)$. This completes the proof of Lemma 9.

Remark 16 Note that if one assumes $||\boldsymbol{Z}_i||_2 \leqslant M$, for all $1 \leqslant i \leqslant N$, and $||\boldsymbol{\theta}||_2 \leqslant \Theta$ (recall the discussion in Remark 2) then, for $\underline{\boldsymbol{\theta}}$ as in (47), by the Cauchy-Schwarz inequality,

$$|\underline{\boldsymbol{\theta}}^\top \boldsymbol{Z}_i| \leqslant \|\underline{\boldsymbol{\theta}}\|_2 \|\boldsymbol{Z}_i\|_2 \leqslant M\left(\|\underline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 + \|\boldsymbol{\theta}\|_2\right) \leqslant M(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 + \Theta).$$

Using this bound and (46) we get,

$$\frac{1}{2N}\sum_{i=1}^{N} \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{U}_i \boldsymbol{U}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\cosh^2(\underline{\beta} m_i(\boldsymbol{X}) + \underline{\boldsymbol{\theta}}^\top \boldsymbol{Z}_i)} \geqslant \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{G} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{2\cosh^2(|\beta| + (M+1)(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1) + M\Theta)}.$$

Note that the bound in the RHS above does not have any dependence on s in the cosh term (unlike in (48)). Hence, by the same arguments as before we can now get the following rate where the dependence on s matches that in the classical high-dimensional logistic regression:

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = O\left(\sqrt{\frac{s \log d}{N}}\right),$$

with probability $1 - o(1)$, as $N, d \to \infty$ such that $d = o(N)$.

A.2.1 Proof of Lemma 14

The first step towards proving Lemma 14 is to observe that:

$$\det(\boldsymbol{G} - \lambda \boldsymbol{I}) = \left(\frac{1}{N}\|\boldsymbol{F}\boldsymbol{m}\|_2^2 - \lambda\right) \cdot \det\left(\frac{1}{N}\boldsymbol{Z}^\top \boldsymbol{Z} - \lambda I\right),$$

where $\boldsymbol{F} := \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$. Hence,

$$\lambda_{\min}(\boldsymbol{G}) = \min\left\{\lambda_{\min}\left(\frac{1}{N}\boldsymbol{Z}^\top \boldsymbol{Z}\right),\ \frac{1}{N}\|\boldsymbol{F}\boldsymbol{m}\|_2^2\right\}.$$

In view of Assumption 2, to prove Lemma 14 it suffices to show that there exists a constant $C > 0$ (depending only on s, Θ and M), such that

$$\mathbb{P}\left(\frac{1}{N}\|\boldsymbol{F}\boldsymbol{m}\|_2^2 \geqslant \frac{C\|\boldsymbol{A}\|_F^2}{N}\right) = 1 - e^{-\Omega(\|\boldsymbol{A}\|_F^4/N)}. \tag{56}$$

To this end, it suffices to prove the following conditional version of (56):

$$\mathbb{P}\left(\frac{1}{N}\|\boldsymbol{F}\boldsymbol{m}\|_2^2 \geqslant \frac{C\|\boldsymbol{A}\|_F^2}{N}\Big|\boldsymbol{X}_{J^c}\right) = 1 - e^{-\Omega(\|\boldsymbol{A}\|_F^4/N)} \tag{57}$$

where J is a suitably chosen subset of $[N]$ and $\boldsymbol{X}_{J^c} := (X_i)_{i \in J^c}$. To this end, note that (3) is a $(|\beta|\|\boldsymbol{A}\|_\infty/2, \Upsilon)$-Ising model (Lemma 12). Now, applying Lemma 11 with

$$\eta := \min\left\{\frac{1}{16},\ \frac{|\beta|\|\boldsymbol{A}\|_\infty}{2}\right\},$$

gives subsets $I_1, \ldots, I_\ell \subseteq [N]$ such that, for all $1 \leqslant j \leqslant \ell$, the conditional distribution of $\boldsymbol{X}_{I_j}$ given $\boldsymbol{X}_{I_j^c} := (X_u)_{u \in [N] \setminus I_j}$ is an (η, Υ)-Ising model. Furthermore, for any vector $\boldsymbol{a}$, there exists $j \in [\ell]$ such that

$$\|\boldsymbol{a}_{I_j}\|_1 \geqslant \frac{\eta}{4|\beta|\|\boldsymbol{A}\|_\infty}\|\boldsymbol{a}\|_1. \tag{58}$$

The proof of (57) now proceeds in the following two steps: First, we show that there exists $j \in [\ell]$ such that the expectation of $N^{-1}\|\boldsymbol{F}\boldsymbol{m}\|_2^2$ conditioned on $\boldsymbol{X}_{I_j}$ is $\Omega(1)$. Subsequently, we establish that conditioned on $\boldsymbol{X}_{I_j}$, $N^{-1}\|\boldsymbol{F}\boldsymbol{m}\|_2^2$ concentrates around its conditional expectation. These steps are verified in Lemmas 17 and 18, respectively.

Lemma 17 *Under the assumptions of Theorem 1, there exists $J \in \{I_1, I_2, \ldots, I_\ell\}$ such that for all $N \geqslant 1$,*

$$\mathbb{E}\left(\frac{1}{N}\|\boldsymbol{F}\boldsymbol{m}\|_2^2 \Big| \boldsymbol{X}_{J^c}\right) \geqslant \frac{C\|\boldsymbol{A}\|_F^2}{N},$$

where $C > 0$ is a constant depending only on Θ, M and s.

Proof For any $(d+1) \times n$ dimensional matrix $\boldsymbol{M}$, we will denote the i-th row of $\boldsymbol{M}$ by $\boldsymbol{M}_i$ and the i-th largest singular value of $\boldsymbol{M}$ by $\sigma_i(\boldsymbol{M})$, for $1 \leqslant i \leqslant d+1$. Also, for $J \subseteq [N]$ denote $(\boldsymbol{F}\boldsymbol{A})_{i,J} := ((\boldsymbol{F}\boldsymbol{A})_{i,j})_{j \in J}$. Note that for any $J \in \{I_1, I_2, \ldots, I_\ell\} \subset [N]$, since $\boldsymbol{m} = \boldsymbol{A}\boldsymbol{X}$,

$$\begin{aligned}
\mathbb{E}\left(\|\boldsymbol{F}\boldsymbol{m}\|_2^2 \Big| \boldsymbol{X}_{J^c}\right) = \sum_{i=1}^N \mathbb{E}\left([(\boldsymbol{F}\boldsymbol{A})_i\boldsymbol{X}]^2 \Big| \boldsymbol{X}_{J^c}\right) &\geqslant \sum_{i=1}^N \mathrm{Var}\left((\boldsymbol{F}\boldsymbol{A})_i\boldsymbol{X} \Big| \boldsymbol{X}_{J^c}\right) \\
&= \sum_{i=1}^N \mathrm{Var}\left((\boldsymbol{F}\boldsymbol{A})_{i,J}\boldsymbol{X}_J \Big| \boldsymbol{X}_{J^c}\right) \\
&\gtrsim \frac{\Upsilon^2}{\eta}\sum_{i=1}^N \|(\boldsymbol{F}\boldsymbol{A})_{i,J}\|_2^2,
\end{aligned} \tag{59}$$

where the last step follows from Lemma 23. Now define a vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_N)^\top$, with $a_i = \|(\boldsymbol{F}\boldsymbol{A})_{\cdot,i}\|_2^2$, where $(\boldsymbol{F}\boldsymbol{A})_{\cdot,i}$ denotes the i-th column of the matrix $\boldsymbol{F}\boldsymbol{A}$. Then by (58) there exists $J \in \{I_1, I_2, \ldots, I_\ell\} \subseteq [N]$ such that

$$\sum_{i=1}^N \|(\boldsymbol{F}\boldsymbol{A})_{i,J}\|_2^2 \geqslant \frac{\eta}{4|\beta|\|\boldsymbol{A}\|_\infty}\sum_{i=1}^N \|(\boldsymbol{F}\boldsymbol{A})_{\cdot,i}\|_2^2.$$

Therefore, by (59),

$$\mathbb{E}\left(\|\boldsymbol{F}\boldsymbol{m}\|_2^2 \Big| \boldsymbol{X}_{J^c}\right) \gtrsim \frac{\Upsilon^2}{|\beta|\|\boldsymbol{A}\|_\infty}\|\boldsymbol{F}\boldsymbol{A}\|_F^2 \gtrsim_{s,B,M} \|\boldsymbol{F}\boldsymbol{A}\|_F^2. \tag{60}$$

By Theorem 2 in Wang and Xi (1997), we get

$$\sum_{t=1}^N \sigma_t^2(\boldsymbol{F}\boldsymbol{A}) \geqslant \sum_{t=1}^N \sigma_t^2(\boldsymbol{F})\ \sigma_{N-t+1}^2(\boldsymbol{A}) \tag{61}$$

Since $\boldsymbol{F}$ is idempotent with trace $N-d$, it follows that $\sigma_1(\boldsymbol{F}) = \ldots = \sigma_{N-d}(\boldsymbol{F}) = 1$ and $\sigma_{N-d+1}(\boldsymbol{F}) = \ldots = \sigma_N(\boldsymbol{F}) = 0$. Hence, we have from (61),

$$\|\boldsymbol{F}\boldsymbol{A}\|_F^2 = \sum_{t=1}^{N} \sigma_t^2(\boldsymbol{F}\boldsymbol{A}) \geqslant \sum_{t=1}^{N-d} \sigma_{N-t+1}^2(\boldsymbol{A}) = \|\boldsymbol{A}\|_F^2 - \sum_{i=1}^{d} \sigma_i^2(\boldsymbol{A}). \tag{62}$$

where the last step uses $\sigma_i^2(\boldsymbol{A}) \leqslant 1$ for all $1 \leqslant i \leqslant N$, since $\|\boldsymbol{A}\|_2 \leqslant \|\boldsymbol{A}\|_\infty \leqslant 1$. Applying the bound in (62) to (60) gives,

$$\mathbb{E}\left(\|\boldsymbol{F}\boldsymbol{m}\|_2^2 \Big| \boldsymbol{X}_{J^c}\right) \gtrsim \Upsilon^2 \left(\|\boldsymbol{A}\|_F^2 - d\right). \tag{63}$$

Lemma 17 now follows from the hypothesis $d = o(\|\boldsymbol{A}\|_F^2)$. ■

Next, we show that $\|\boldsymbol{F}\boldsymbol{m}\|_2^2$ concentrates around its conditional expectation $\mathbb{E}(\|\boldsymbol{F}\boldsymbol{m}\|_2^2|\boldsymbol{X}_{J^c})$, for the set J as defined above.

Lemma 18 *For any $t > 0$ and $J \in \{I_1, I_2, \ldots, I_\ell\}$,*

$$\begin{aligned}&\mathbb{P}\left(\|\boldsymbol{F}\boldsymbol{m}\|_2^2 < \mathbb{E}(\|\boldsymbol{F}\boldsymbol{m}\|_2^2|\boldsymbol{X}_{J^c}) - t \Big| \boldsymbol{X}_{J^c}\right) \\ &\leqslant 2\exp\left(-C \cdot \min\left\{\frac{t^2}{8N}, \frac{t}{2}\right\}\right) + 2\exp\left(-\frac{t^2}{128N}\right),\end{aligned} \tag{64}$$

where C is a constant depending only on Θ, M and s.

Proof Denote $\boldsymbol{W} := \boldsymbol{F}\boldsymbol{A}$ and let $\boldsymbol{H}$ be the matrix obtained from $\boldsymbol{W}^\top\boldsymbol{W}$ by zeroing out all its diagonal elements. Moreover, throughout the proof we will denote $I := J^c$. Clearly,

$$\|\boldsymbol{F}\boldsymbol{m}\|_2^2 - \mathbb{E}(\|\boldsymbol{F}\boldsymbol{m}\|_2^2|\boldsymbol{X}_I) = \boldsymbol{X}^\top\boldsymbol{H}\boldsymbol{X} - \mathbb{E}\left(\boldsymbol{X}^\top\boldsymbol{H}\boldsymbol{X}|\boldsymbol{X}_I\right).$$

By permuting the indices let us partition the vector $\boldsymbol{X}$ as $(\boldsymbol{X}_I^\top, \boldsymbol{X}_J^\top)^\top$ and the matrix $\boldsymbol{H}$ as:

$$\begin{pmatrix} \boldsymbol{H}_{I,I} & \boldsymbol{H}_{I,J} \\ \boldsymbol{H}_{I,J}^\top & \boldsymbol{H}_{J,J} \end{pmatrix}$$

where for two subsets A, B of $[N]$, we define $\boldsymbol{H}_{A,B} := ((H_{ij}))_{i\in A, j\in B}$. Note that

$$\begin{aligned}&\boldsymbol{X}^\top\boldsymbol{H}\boldsymbol{X} - \mathbb{E}\left(\boldsymbol{X}^\top\boldsymbol{H}\boldsymbol{X}|\boldsymbol{X}_I\right) \\ &= \boldsymbol{X}_J^\top\boldsymbol{H}_{J,J}\boldsymbol{X}_J - \mathbb{E}\left(\boldsymbol{X}_J^\top\boldsymbol{H}_{J,J}\boldsymbol{X}_J|\boldsymbol{X}_I\right) + 2\boldsymbol{X}_I^\top\boldsymbol{H}_{I,J}\boldsymbol{X}_J - 2\mathbb{E}\left(\boldsymbol{X}_I^\top\boldsymbol{H}_{I,J}\boldsymbol{X}_J|\boldsymbol{X}_I\right).\end{aligned} \tag{65}$$

Since $\boldsymbol{X}|\boldsymbol{X}_I$ is an (η, Υ)-Ising model, Example 2.5 in Adamczak et al. (2019) implies (by taking the parameters α and ρ in Adamczak et al. (2019) to be $\Theta M s + 1/16$ and $7/8$, respectively),

$$\begin{aligned}&\mathbb{P}\left(\boldsymbol{X}_J^\top\boldsymbol{H}_{J,J}\boldsymbol{X}_J < \mathbb{E}\left(\boldsymbol{X}_J^\top\boldsymbol{H}_{J,J}\boldsymbol{X}_J|\boldsymbol{X}_I\right) - t \Big| \boldsymbol{X}_I\right) \\ &\leqslant 2\exp\left(-c \cdot \min\left\{\frac{t^2}{\|\boldsymbol{H}_{J,J}\|_F^2 + \|\mathbb{E}(\boldsymbol{H}_{J,J}\boldsymbol{X}_J)\|_2^2}, \frac{t}{\|\boldsymbol{H}_{J,J}\|_2}\right\}\right),\end{aligned} \tag{66}$$

where c is a constant depending only on Θ, M and s. Clearly, one has $\|\boldsymbol{H}_{J,J}\|_F^2 \leqslant \|\boldsymbol{H}\|_F^2$ and, since the spectral norm of a matrix is always greater than or equal to the spectral norm of any of its submatrices, $\|\boldsymbol{H}_{J,J}\|_2 \leqslant \|\boldsymbol{H}\|_2$. Moreover, if

$$\boldsymbol{X}_0^\top := (\boldsymbol{0}^\top, \boldsymbol{X}_J^\top),$$

then $\boldsymbol{H}_{J,J}\boldsymbol{X}_J$ is a subvector of $\boldsymbol{H}\boldsymbol{X}_0$, and hence,

$$\|\mathbb{E}(\boldsymbol{H}_{J,J}\boldsymbol{X}_J)\|_2^2 \leqslant \|\mathbb{E}(\boldsymbol{H}\boldsymbol{X}_0)\|_2^2.$$

Combining all these, we have from (66)

$$\begin{aligned}&\mathbb{P}\left(\boldsymbol{X}_J^\top \boldsymbol{H}_{J,J}\boldsymbol{X}_J < \mathbb{E}\left(\boldsymbol{X}_J^\top \boldsymbol{H}_{J,J}\boldsymbol{X}_J|\boldsymbol{X}_I\right) - t\Big|\boldsymbol{X}_I\right)\\ &\leqslant 2\exp\left(-c\cdot\min\left\{\frac{t^2}{\|\boldsymbol{H}\|_F^2 + \|\mathbb{E}(\boldsymbol{H}\boldsymbol{X}_0)\|_2^2},\ \frac{t}{\|\boldsymbol{H}\|_2}\right\}\right)\end{aligned} \tag{67}$$

Next, note that, since $\boldsymbol{H} - \boldsymbol{W}^\top\boldsymbol{W}$ is a diagonal matrix,

$$\begin{aligned}\|\boldsymbol{H}\|_F^2 + \|\mathbb{E}(\boldsymbol{H}\boldsymbol{X}_0)\|_2^2 &\leqslant \|\boldsymbol{H}\|_F^2 + \left(\|\mathbb{E}(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{X}_0)\|_2 + \|\mathbb{E}[(\boldsymbol{H} - \boldsymbol{W}^\top\boldsymbol{W})\boldsymbol{X}_0]\|_2\right)^2\\ &\leqslant 2\|\boldsymbol{H}\|_F^2 + 2\mathbb{E}\|[(\boldsymbol{H} - \boldsymbol{W}^\top\boldsymbol{W})\boldsymbol{X}_0]\|_2^2 + 2\|\mathbb{E}(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{X}_0)\|_2^2\\ &= 2\sum_{1\leqslant i\neq j\leqslant N}(\boldsymbol{W}^\top\boldsymbol{W})_{ij}^2 + 2\sum_{i=1}^N(\boldsymbol{W}^\top\boldsymbol{W})_{ii}^2 + 2\|\mathbb{E}(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{X}_0)\|_2^2\\ &= 2\|\boldsymbol{W}^\top\boldsymbol{W}\|_F^2 + 2\|\mathbb{E}(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{X}_0)\|_2^2.\end{aligned} \tag{68}$$

We also have $\|\boldsymbol{H}\|_2 \leqslant \|\boldsymbol{W}^\top\boldsymbol{W}\|_2$, since for any vector $\boldsymbol{a} \in \mathbb{R}^N$,

$$\boldsymbol{a}^\top\boldsymbol{H}\boldsymbol{a} = \boldsymbol{a}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{a} - \sum_{i=1}^N(\boldsymbol{W}^\top\boldsymbol{W})_{ii}a_i^2 \leqslant \boldsymbol{a}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{a}.$$

Hence, it follows from (67) and (68) that

$$\begin{aligned}&\mathbb{P}\left(\boldsymbol{X}_J^\top \boldsymbol{H}_{J,J}\boldsymbol{X}_J < \mathbb{E}\left(\boldsymbol{X}_J^\top \boldsymbol{H}_{J,J}\boldsymbol{X}_J|\boldsymbol{X}_I\right) - t\Big|\boldsymbol{X}_I\right)\\ &\leqslant 2\exp\left(-c'\cdot\min\left\{\frac{t^2}{\|\boldsymbol{W}^\top\boldsymbol{W}\|_F^2 + \|\mathbb{E}(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{X}_0)\|_2^2},\ \frac{t}{\|\boldsymbol{W}^\top\boldsymbol{W}\|_2}\right\}\right)\end{aligned} \tag{69}$$

where $c' := c/2$. Next, recall that for any two matrices $\boldsymbol{U}$ and $(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ such that $\boldsymbol{U}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ exists, we have $\|\boldsymbol{U}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_F \leqslant \|\boldsymbol{U}\|_2\ \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_F$. This implies,

$$\|\boldsymbol{W}^\top\boldsymbol{W}\|_F^2 \leqslant \|\boldsymbol{W}\|_2^2\ \|\boldsymbol{W}\|_F^2, \tag{70}$$

Moreover, by the submultiplicativity of the matrix ℓ_2 norm,

$$\|\boldsymbol{W}^\top\boldsymbol{W}\|_2 \leqslant \|\boldsymbol{W}^\top\|_2\|\boldsymbol{W}\|_2 = \|\boldsymbol{W}\|_2^2 \tag{71}$$

and

$$\|\mathbb{E}(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{X}_0)\|_2^2 = \|\boldsymbol{W}^\top\mathbb{E}(\boldsymbol{W}\boldsymbol{X}_0)\|_2^2 \leqslant \|\boldsymbol{W}\|_2^2\cdot\|\mathbb{E}(\boldsymbol{W}\boldsymbol{X}_0)\|_2^2. \tag{72}$$

It follows from (69), (70), (71) and (72), that:

$$\mathbb{P}\left(\boldsymbol{X}_J^\top \boldsymbol{H}_{J,J}\boldsymbol{X}_J < \mathbb{E}\left(\boldsymbol{X}_J^\top \boldsymbol{H}_{J,J}\boldsymbol{X}_J|\boldsymbol{X}_I\right) - t\Big|\boldsymbol{X}_I\right) \\ \leqslant 2\exp\left(-\frac{c'}{\|\boldsymbol{F}\boldsymbol{A}\|_2^2}\cdot\min\left\{\frac{t^2}{\|\boldsymbol{F}\boldsymbol{A}\|_F^2+\|\mathbb{E}\left[\boldsymbol{W}\boldsymbol{X}_0\right]\|_2^2},\ t\right\}\right). \tag{73}$$

Since $\|\boldsymbol{F}\boldsymbol{A}\|_2^2 \leqslant \|\boldsymbol{F}\|_2^2\|\boldsymbol{A}\|_2^2 \leqslant 1$, $\|\boldsymbol{F}\boldsymbol{A}\|_F^2 \leqslant N\|\boldsymbol{F}\boldsymbol{A}\|_2^2 \leqslant N$, and $\|\mathbb{E}\left[\boldsymbol{W}\boldsymbol{X}_0\right]\|_2^2 \leqslant \mathbb{E}\|\boldsymbol{W}\boldsymbol{X}_0\|_2^2 \leqslant \|\boldsymbol{F}\boldsymbol{A}\|_2^2\,\mathbb{E}\|\boldsymbol{X}_0\|_2^2 \leqslant N$, we can conclude from (73) that:

$$\mathbb{P}\left(\boldsymbol{X}_J^\top \boldsymbol{H}_{J,J}\boldsymbol{X}_J < \mathbb{E}\left(\boldsymbol{X}_J^\top \boldsymbol{H}_{J,J}\boldsymbol{X}_J|\boldsymbol{X}_I\right) - t\Big|\boldsymbol{X}_I\right) \leqslant 2\exp\left(-c'\cdot\min\left\{\frac{t^2}{2N},\ t\right\}\right). \tag{74}$$

Next, let us define $\boldsymbol{y} := \boldsymbol{H}_{I,J}^\top \boldsymbol{X}_I$. Then,

$$\boldsymbol{X}_I^\top \boldsymbol{H}_{I,J}\boldsymbol{X}_J - \mathbb{E}\left(\boldsymbol{X}_I^\top \boldsymbol{H}_{I,J}\boldsymbol{X}_J|\boldsymbol{X}_I\right) = \boldsymbol{y}^\top \boldsymbol{X}_J - \mathbb{E}(\boldsymbol{y}^\top \boldsymbol{X}_J|\boldsymbol{X}_I).$$

By Lemma 4.4 in Chatterjee (2016), Dobrushin's interdependence matrix for the model $\boldsymbol{X}_J|\boldsymbol{X}_I$ is given by $8\boldsymbol{D}$, where $\boldsymbol{D}$ denotes the interaction matrix for the Ising model $\boldsymbol{X}_J|\boldsymbol{X}_I$. Since $\|8\boldsymbol{D}\|_2 \leqslant \frac{1}{2}$, by Theorem 4.3 in Chatterjee (2016),

$$\mathbb{P}\left(\boldsymbol{y}^\top \boldsymbol{X}_J < \mathbb{E}(\boldsymbol{y}^\top \boldsymbol{X}_J|\boldsymbol{X}_I) - t\ \Big|\boldsymbol{X}_I\right) \leqslant 2\exp\left(-\frac{t^2}{8\|\boldsymbol{y}\|_2^2}\right). \tag{75}$$

Using the submultiplicativity of the matrix ℓ_2 norm and the fact that the spectral norm of a matrix is always greater than or equal to the spectral norm of any of its submatrices gives,

$$\|\boldsymbol{y}\|_2^2 = \|\boldsymbol{H}_{I,J}^\top \boldsymbol{X}_I\|_2^2 \leqslant \|\boldsymbol{H}_{I,J}^\top\|_2^2\cdot\|\boldsymbol{X}_I\|_2^2 \leqslant N\|\boldsymbol{H}\|_2^2 \leqslant N\|\boldsymbol{W}\|_2^4 = N\|\boldsymbol{F}\boldsymbol{A}\|_2^4 \leqslant N. \tag{76}$$

Combining (75) and (76),

$$\mathbb{P}\left(\boldsymbol{y}^\top \boldsymbol{X}_J < \mathbb{E}(\boldsymbol{y}^\top \boldsymbol{X}_J|\boldsymbol{X}_I) - t\ \Big|\boldsymbol{X}_I\right) \leqslant 2\exp\left(-\frac{t^2}{8N}\right). \tag{77}$$

Finally, combining (74), (77), and (65) gives,

$$\mathbb{P}\left(\boldsymbol{X}^\top \boldsymbol{H}\boldsymbol{X} < \mathbb{E}\left(\boldsymbol{X}^\top \boldsymbol{H}\boldsymbol{X}|\boldsymbol{X}_I\right) - t\ \Big|\boldsymbol{X}_I\right) \\ \leqslant 2\exp\left(-c'\cdot\min\left\{\frac{t^2}{8N},\ \frac{t}{2}\right\}\right) + 2\exp\left(-\frac{t^2}{128N}\right).$$

This completes the proof of Lemma 18. ■

To complete the proof of (57), we choose J as in Lemma 17 and apply Lemma 18 with $t = \frac{1}{2}\mathbb{E}(\|\boldsymbol{F}\boldsymbol{m}\|_2^2|\boldsymbol{X}_{J^c})$. This implies, there exists a constant C, depending only on Θ, M and s, such that

$$\mathbb{P}\left(\frac{1}{N}\|\boldsymbol{F}\boldsymbol{m}\|_2^2 \geqslant \frac{C\|\boldsymbol{A}\|_F^2}{N}\Big|\boldsymbol{X}_{J^c}\right) = 1 - e^{-\Omega(\|\boldsymbol{A}\|_F^4/N)}.$$

This proves (57) and completes the proof of Lemma 14. □

Appendix B. Proof of Theorem 2

Recall the definition of the log-pseudo-likelihood function $L_{\beta,N}(\cdot)$ from (10). As before, the proof of Theorem 1 entails showing the following: (1) concentration of the gradient of $L_{\beta,N}$ and (2) restricted strong concavity of $L_{\beta,N}$.

The concentration of the gradient $\nabla L_{\beta,N}$ follows by arguments similar to Lemma 8. Towards this, recall the definition of the functions $\phi_{i,s} : \mathcal{C}_N \to \mathbb{R}$, for $1 \leqslant i \leqslant N$ and $1 \leqslant s \leqslant d$, from (31). Note that $\nabla L_{\beta,N} = (\frac{\partial L_{\beta,N}}{\partial \theta_1}, \ldots, \frac{\partial L_{\beta,N}}{\partial \theta_d})^\top$, where $\frac{\partial L_{\beta,N}}{\partial \theta_s} = \sum_{i=1}^N \phi_{i,s}(\boldsymbol{X})$ for $1 \leqslant s \leqslant d$. For $k \in [N]$, recall from Lemma 13 the definition of

$$c_k := \frac{2|Z_{k,s}|}{N} + \frac{2|\beta|}{N} \sum_{i=1}^N |Z_{i,s} a_{ik}|\,,$$

where $s \in [d]$. Therefore, for $s \in [d]$,

$$\begin{aligned}
\sum_{k=1}^N c_k^2 &\lesssim \frac{1}{N^2} \sum_{k=1}^N Z_{k,s}^2 + \frac{\beta^2}{N^2} \sum_{k=1}^N \left(\sum_{i=1}^N |Z_{i,s} a_{ik}| \right)^2 \\
&\leqslant \frac{1}{N} \max_{1 \leqslant s \leqslant d} \frac{1}{N} \sum_{k=1}^N Z_{k,s}^2 + \frac{\beta^2}{N^2} \sum_{k=1}^N \left(\sum_{i=1}^N Z_{i,s}^2 |a_{ik}| \sum_{i=1}^N |a_{ik}| \right) \\
&\qquad\qquad \text{(by the Cauchy-Schwarz inequality)} \\
&\leqslant \frac{1}{N} \max_{1 \leqslant s \leqslant d} \frac{1}{N} \sum_{k=1}^N Z_{k,s}^2 + \left(\max_{1 \leqslant k \leqslant N} \sum_{i=1}^N |a_{ik}| \right) \frac{\beta^2}{N^2} \sum_{i=1}^N Z_{i,s}^2 \sum_{k=1}^N |a_{ik}| \\
&\leqslant \frac{1}{N} \left\{ \max_{1 \leqslant s \leqslant d} \frac{1}{N} \sum_{k=1}^N Z_{k,s}^2 + \beta^2 \|\boldsymbol{A}\|_1^2 \max_{1 \leqslant s \leqslant d} \frac{1}{N} \sum_{i=1}^N Z_{i,s}^2 \right\} \\
&\lesssim_{\beta, C} \frac{1}{N},
\end{aligned}$$

where the last holds with probability 1 under Assumptions 1 and 4. Then by analogous arguments as in Lemma 8 it follows that there exists $\delta > 0$ such that with $\lambda := \delta \sqrt{\log d / N}$ the following holds:

$$\mathbb{P}\left(\|\nabla L_{\beta,N}(\boldsymbol{\theta})\|_\infty > \frac{\lambda}{2} \right) = o(1), \tag{78}$$

where the $o(1)$-term goes to infinity as $d \to \infty$.

To establish the strong concavity of $L_{\beta,N}$ we consider the second-order Taylor expansion: For any $\boldsymbol{\eta} \in \mathbb{R}^d$,

$$L_{\beta,N}(\boldsymbol{\theta} + \boldsymbol{\eta}) - L_{\beta,N}(\boldsymbol{\theta}) - \nabla L_{\beta,N}(\boldsymbol{\theta})^\top \boldsymbol{\eta} = \frac{1}{2} \boldsymbol{\eta}^\top \nabla^2 L_{\beta,N}(\boldsymbol{\theta} + t\boldsymbol{\eta}) \boldsymbol{\eta},$$

for some $t \in (0,1)$. Computing the Hessian matrix $\nabla^2 L_{\beta,N}(\boldsymbol{\theta} + t\boldsymbol{\eta})$ gives,

$$\begin{aligned}
L_{\beta,N}(\boldsymbol{\theta}+\boldsymbol{\eta}) - L_{\beta,N}(\boldsymbol{\theta}) - \nabla L_{\beta,N}(\boldsymbol{\theta})^\top \boldsymbol{\eta} & \\
&= \frac{1}{2N}\sum_{i=1}^{N} \frac{\boldsymbol{\eta}^\top \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \boldsymbol{\eta}}{\cosh^2(\beta m_i(\boldsymbol{X}) + (\boldsymbol{\theta}+t\boldsymbol{\eta})^\top \boldsymbol{Z}_i)} \\
&\geqslant \frac{1}{2N}\sum_{i=1}^{N} \frac{\boldsymbol{\eta}^\top \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \boldsymbol{\eta}}{\cosh^2(|\beta|\|\boldsymbol{A}\|_\infty + (\boldsymbol{\theta}+t\boldsymbol{\eta})^\top \boldsymbol{Z}_i)} \\
&\qquad\qquad \text{(using } |\beta m_i(\boldsymbol{X})| \leqslant |\beta|||\boldsymbol{A}||_\infty) \\
&= \frac{1}{2N}\sum_{i=1}^{N} \langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle^2 \frac{1}{\cosh^2(|\beta|\|\boldsymbol{A}\|_\infty + (\boldsymbol{\theta}+t\boldsymbol{\eta})^\top \boldsymbol{Z}_i)} \\
&= \frac{1}{2N}\sum_{i=1}^{N} \langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle^2 \psi(\langle \boldsymbol{\theta}, \boldsymbol{Z}_i\rangle + t\langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle), \qquad (79)
\end{aligned}$$

where $\psi(x) := \operatorname{sech}^2(|\beta|\|\boldsymbol{A}\|_\infty + x)$.

Proposition 19 *Suppose the assumptions of Theorem 2 hold. Then there exists positive constants ν, c_0 (depending on κ_1, κ_2, β, and Θ) such that for all $t \in (0,1)$,*

$$\frac{1}{N}\sum_{i=1}^{N} \langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle^2 \psi(\langle \boldsymbol{\theta}, \boldsymbol{Z}_i\rangle + t\langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle) \geqslant \nu\|\boldsymbol{\eta}\|_2^2 - c_0 \mathcal{R}_N \|\boldsymbol{\eta}\|_1^2,$$

with probability at least $1 - o(1)$, for all $\boldsymbol{\eta} \in \mathbb{R}^d$ with $\|\boldsymbol{\eta}\|_2 \leqslant 1$.

The proof of Proposition 19 is given in Section B.1. Here we use it to complete the proof of Theorem 2. Note that by (79) and Proposition 19, for any $\boldsymbol{\eta} \in \mathbb{R}^d$ with $\|\boldsymbol{\eta}\|_2 \leqslant 1$,

$$L_{\beta,N}(\boldsymbol{\theta}+\boldsymbol{\eta}) - L_{\beta,N}(\boldsymbol{\theta}) - \nabla L_{\beta,N}(\boldsymbol{\theta})^\top \boldsymbol{\eta} \gtrsim \nu\|\boldsymbol{\eta}\|_2^2 - c_0 \mathcal{R}_N \|\boldsymbol{\eta}\|_1^2$$

This establishes the restricted strong convexity (RSC) property for pseudo-likelihood loss function $L_{\beta,N}(\cdot)$. Therefore, by Wainwright (2019, Corollary 9.20), whenever $\mathcal{R}_N s = s\sqrt{\log d/n} = o(1)$, then when the event $\{\|\nabla L_N(\hat{\boldsymbol{\theta}})\|_\infty \leqslant \frac{\lambda}{2}\}$ happens,

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \lesssim_\nu s\lambda^2 \lesssim_{\nu,\delta} \frac{s\log d}{n} \quad \text{and} \quad \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 \lesssim_\nu s\lambda \lesssim_{\nu,\delta} s\sqrt{\frac{\log d}{n}}. \qquad (80)$$

Since the event $\{\|\nabla L_N(\hat{\boldsymbol{\theta}})\|_\infty \leqslant \frac{\lambda}{2}\}$ happens with probability $1 - o(1)$ by (78) (jointly over the randomness of the data and the covariates), the bounds in (80) hold with probability $1 - o(1)$, which completes the proof of Theorem 2.

B.1 Proof of Proposition 19

Consider a fixed vector $\boldsymbol{\eta} \in \mathbb{R}^d$ with $\|\boldsymbol{\eta}\|_2 = r \in (0,1]$ and set $L = L(r) := Kr$, for a constant $K > 0$ to be chosen. Since the function $\phi_L(u) := u^2 I\{|u| \leqslant 2L\} \leqslant u^2$ and ψ is

positive,

$$\frac{1}{N}\sum_{i=1}^{N}\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle^2\psi(\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle+t\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle)\geqslant\frac{1}{N}\sum_{i=1}^{N}\psi(\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle+t\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle)\phi_L(\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle)\mathbf{1}\left\{|\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle|\leqslant T\right\},$$

where T is second truncation parameter to be chosen. Note that on the events $\{|\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle|\leqslant 2L\}$ and $\{|\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle|\leqslant T\}$ one has,

$$|\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle+t\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle|\leqslant T+2L\leqslant T+2K,$$

since $t,r\in[0,1]$. This implies,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle^2\psi(\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle+t\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle)\\
&\quad\geqslant\frac{1}{N}\sum_{i=1}^{N}\psi(\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle+t\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle)\phi_L(\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle)\mathbf{1}\left\{|\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle|\leqslant T\right\}\\
&\quad\geqslant\frac{\gamma}{N}\sum_{i=1}^{N}\phi_L(\langle\boldsymbol{\eta},Z_i\rangle)\mathbf{1}\left\{|\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle|\leqslant T\right\},
\end{aligned}\tag{81}$$

where $\gamma:=\min_{|u|\leqslant T+2K}\psi(u)>0$. Based on this lower bound, to prove Proposition 19 it is sufficient to show that for all $r\in(0,1]$ and for $\boldsymbol{\eta}\in\mathbb{R}^d$ with $\|\boldsymbol{\eta}\|_2=r$,

$$\frac{1}{N}\sum_{i=1}^{N}\phi_{L(r)}(\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle)\mathbf{1}\left\{|\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle|\leqslant T\right\}\geqslant c_1r^2-c_2\mathcal{R}_N\|\boldsymbol{\eta}\|_1 r,\tag{82}$$

holds with probability $1-o(1)$, where $L(r)=Kr$ and some positive constants c_1,c_2. This is because when (82) holds by substituting $\|\boldsymbol{\eta}\|_2=r$ and using (81) gives,

$$\frac{1}{N}\sum_{i=1}^{N}\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle^2\psi(\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle+t\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle)\geqslant\nu\|\boldsymbol{\eta}\|_2^2-c_0\mathcal{R}_N\|\boldsymbol{\eta}\|_1^2,$$

where the constants (ν,c_0) depend on (c_1,c_2,γ) and we use the inequality $\|\boldsymbol{\eta}\|_2\leqslant\|\boldsymbol{\eta}\|_1$. In fact, it suffices to prove (82) for $r=1$, that is,

$$\frac{1}{N}\sum_{i=1}^{N}\phi_K(\langle\boldsymbol{\eta},\boldsymbol{Z}_i\rangle)\mathbf{1}\left\{|\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle|\leqslant T\right\}\geqslant c_1-c_2\mathcal{R}_N\|\boldsymbol{\eta}\|_1,\tag{83}$$

holds with probability $1-o(1)$. This is because given any vector in $\mathbb{R}^d$ with $\|\boldsymbol{\eta}\|_2=r>0$, we can apply (83) to the rescaled unit-norm vector $\boldsymbol{\eta}/r$ to obtain

$$\frac{1}{N}\sum_{i=1}^{N}\phi_K(\langle\boldsymbol{\eta}/r,\boldsymbol{Z}_i\rangle)\mathbf{1}\left\{|\langle\boldsymbol{\theta},\boldsymbol{Z}_i\rangle|\leqslant T\right\}\geqslant c_1-c_2\mathcal{R}_N\frac{\|\boldsymbol{\eta}\|_1}{r}.\tag{84}$$

Noting that $\phi_K(u/r)=\phi_{L(1)}(u/r)=(1/r)^2\phi_{L(r)}(u)$ and multiplying both sides of (84) by r^2 gives (82).

B.1.1 Proof of (83)

Define a new truncation function:

$$\bar{\phi}_K(u) = u^2\mathbf{1}\left\{|u| \leqslant K\right\} + (u-2K)^2\mathbf{1}\left\{K < u \leqslant 2K\right\} + (u+2K)^2\mathbf{1}\left\{-2K \leqslant u < -K\right\}.$$

Note this function is Lipschitz with parameter $2K$. Moreover, since $\phi_K \geqslant \bar{\phi}_K$, to prove (83) it suffices to show that for all vectors $\boldsymbol{\eta} \in \mathbb{R}^d$ of unit norm,

$$\frac{1}{N}\sum_{i=1}^{N} \bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}_i\rangle| \leqslant T\right\} \geqslant c_1 - c_2\mathcal{R}_N\|\boldsymbol{\eta}\|_1, \tag{85}$$

holds with probability $1 - o(1)$. To this end, for a given ℓ_1 radius $b \geqslant 1$, define the random variable

$$Y_n(b) := \sup_{\substack{\|\boldsymbol{\eta}\|_2 = 1 \\ \frac{b}{2} \leqslant \|\boldsymbol{\eta}\|_1 \leqslant b}} \left| \frac{1}{N}\sum_{i=1}^{N} \bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}_i\rangle| \leqslant T\right\} - \mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| \leqslant T\right\}\right) \right|.$$

Lemma 20 *Under the assumptions of Theorem 2 the following hold:*

(1) *By choosing* $K^2 = 8\kappa_2/\kappa_1$ *and* $T^2 = 8\kappa_2\Theta^2/\kappa_1$,

$$\mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| \leqslant T\right\}\right) \geqslant \frac{3}{4}\kappa_1.$$

(2) *There exist a positive constant* c_2 *such that*

$$\mathbb{P}\left(Y_n(b) > \frac{1}{2}\kappa_1 + \frac{1}{2}c_2\mathcal{R}_N b\right) \leqslant e^{-O_{\kappa_1,\kappa_2}(n)}.$$

Lemma 20 implies that with probability $1 - e^{-O_{\kappa_1,\kappa_2}(n)}$ for all $\boldsymbol{\eta} \in \mathbb{R}^d$ such that $\|\boldsymbol{\eta}\|_2 = 1$ and $\frac{b}{2} \leqslant \|\boldsymbol{\eta}\|_1 \leqslant b$, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}_i\rangle| \leqslant T\right\} &\geqslant \mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| \leqslant T\right\}\right) - \frac{1}{2}\kappa_1 - \frac{1}{2}c_2\mathcal{R}_N b \\ &\geqslant \frac{1}{4}\kappa_1 - c_2\mathcal{R}_N\|\boldsymbol{\eta}\|_1. \end{aligned} \tag{86}$$

This establishes the bound (85) with $c_1 = \kappa_1/4$ for all vectors $\boldsymbol{\eta} \in \mathbb{R}^d$ with $\|\boldsymbol{\eta}\|_2 = 1$ and $\frac{b}{2} \leqslant \|\boldsymbol{\eta}\|_1 \leqslant b$, with probability $1 - e^{-O_{\kappa_1,\kappa_2}(n)}$.

We first prove Lemma 20. Then we will extend the bound in (85) for all vectors $\boldsymbol{\eta} \in \mathbb{R}^d$ with $\|\boldsymbol{\eta}\|_2 = 1$ using a peeling strategy to complete the proof.

Proof of Lemma 20 (1)

To show Lemma 20 (1) it suffices to show that

$$\mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\right) \geqslant \frac{7}{8}\kappa_1 \quad \text{and} \quad \mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| > T\right\}\right) \leqslant \frac{1}{8}\kappa_1. \tag{87}$$

Indeed, if these two inequalities hold, then

$$\begin{aligned}\mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| \leqslant T\right\}\right) &= \mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\right) - \mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| > T\right\}\right) \\ &\geqslant \frac{7}{8}\kappa_1 - \frac{1}{8}\kappa_2 = \frac{3}{4}\kappa_1.\end{aligned}$$

To prove first inequality in (87), note that

$$\begin{aligned}\mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\right) \geqslant \mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^2\mathbf{1}\left\{|\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle| \leqslant K\right\}\right) &= \mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^2\right) - \mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^2\mathbf{1}\left\{|\boldsymbol{\eta}, \boldsymbol{Z}| > K\right\}\right) \\ &\geqslant \kappa_1 - \mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^2\mathbf{1}\left\{|\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle| > K\right\}\right). \end{aligned} \tag{88}$$

To lower bound the second term, applying the Cauchy-Schwarz inequality yields,

$$\begin{aligned}\mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^2\mathbf{1}\left\{|\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle| > K\right\}\right) &\leqslant \sqrt{\mathbb{E}\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^4\right)}\sqrt{\mathbb{P}\left(|\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle| > K\right)} \\ &\leqslant \frac{\mathbb{E}(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^4)}{K^2} \qquad \text{(by Markov's inequality)} \\ &\leqslant \frac{\kappa_2}{K^2},\end{aligned}$$

using the assumption $\mathbb{E}(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^4) \leqslant \kappa_2$, for all $\boldsymbol{\eta}$ such that $\|\boldsymbol{\eta}\|_2 \leqslant 1$. Therefore, setting $K^2 = 8\kappa_2/\kappa_1$ and using (88) proves the first inequality in (87).

Now, we turn to prove the second inequality in (87). For this note that $\bar{\phi}_K\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle\right) \leqslant \langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^2$ and by Markov's inequality,

$$\mathbb{P}\left(|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| \geqslant T\right) \leqslant \frac{\kappa_2\|\boldsymbol{\theta}\|_2^4}{T^4}.$$

Then the Cauchy-Schwarz inequality implies,

$$\mathbb{E}\left(\bar{\phi}_K\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle\right)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| > T\right\}\right) \leqslant \frac{\kappa_2\|\boldsymbol{\theta}\|_2^2}{T^2} \leqslant \frac{\kappa_2\Theta^2}{T^2}$$

Thus setting $T^2 = 8\kappa_2\Theta^2/\kappa_1$ shows the second inequality in (87).

Proof of Lemma 20 (2)

We begin by recalling the functional Hoeffding's inequality. Towards this, let $\mathcal{F}$ be a symmetric collection of functions from $\mathbb{R}^d$ to $\mathbb{R}$, that is, if $f \in \mathcal{F}$, then $-f \in \mathcal{F}$. Suppose $X_1, X_2, \ldots, X_N$ are i.i.d. from a distribution supported on $\mathcal{X} \subseteq \mathbb{R}^d$ and let

$$Z := \sup_{f\in\mathcal{F}}\left\{\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right\}.$$

Then we have the following result:

Lemma 21 ((Wainwright, 2019, Theorem 3.26)) *For each $f \in \mathcal{F}$ assume that there are real numbers $a_f \leqslant b_f$ such that $f(x) \in [a_f, b_f]$ for all $x \in \mathcal{X}$. Then for all $\delta \geqslant 0$,*

$$\mathbb{P}\left(Z - \mathbb{E}(Z) \geqslant \delta\right) \leqslant \exp\left(-\frac{n\delta^2}{4L^2}\right)$$

where $L^2 := \sup_{f\in\mathcal{F}} \left\{(b_f - a_f)^2\right\}$.

For $\boldsymbol{\eta} \in \mathbb{R}^d$ such that $\|\boldsymbol{\eta}\|_2 = 1$ and $\|\boldsymbol{\eta}\|_1 \leqslant b$, define

$$f_{\boldsymbol{\eta}}(\boldsymbol{Z}) := \bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| \leqslant T\right\} - \mathbb{E}\left(\bar{\phi}_K(\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}\rangle| \leqslant T\right\}\right)$$

Note that

$$Y_n(b) = \sup_{\substack{\|\boldsymbol{\eta}\|_2=1 \\ \frac{b}{2}\leqslant\|\boldsymbol{\eta}\|_1\leqslant b}} \left\{\left|\frac{1}{N}\sum_{i=1}^N f_{\boldsymbol{\eta}}(\boldsymbol{Z}_i)\right|\right\}.$$

Since $|f_{\boldsymbol{\eta}}(\boldsymbol{z})| \leqslant K^2$, for any positive constant c_3 by Lemma 21,

$$\mathbb{P}\left(Y_n(b) - \mathbb{E}(Y_n(b)) \geqslant c_3\mathcal{R}_N b + \frac{1}{2}\kappa_1\right) \leqslant e^{-c_4 n\mathcal{R}_N^2 b^2 - c_4 n} \leqslant e^{-c_4 n}, \tag{89}$$

where c_4 is a positive constant depending on K and c_3. Now, suppose $\{\varepsilon_i\}_{i=1}^n$ be a sequence of i.i.d. Rademacher variables. Then a symmetrization argument (Wainwright, 2019, Proposition 4.11) implies that

$$\mathbb{E}\left(Y_n(b)\right) \leqslant 2\mathbb{E}_{\boldsymbol{Z},\boldsymbol{\varepsilon}}\left(\sup_{\substack{\|\boldsymbol{\eta}\|_2=1 \\ \frac{b}{2}\leqslant\|\boldsymbol{\eta}\|_1\leqslant b}} \left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i \bar{\phi}_K\left(\langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle\right)\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}_i\rangle| \leqslant T\right\}\right|\right).$$

Since $\mathbf{1}\left\{|\langle \boldsymbol{\theta}, \boldsymbol{Z}_i\rangle| \leqslant T\right\} \leqslant 1$ and $\bar{\phi}_K$ is Lipschitz with parameter $2K$, the contraction principle yields

$$\begin{aligned}
\mathbb{E}\left(Y_n(b)\right) &\leqslant 8K\mathbb{E}_{\boldsymbol{Z},\boldsymbol{\varepsilon}}\left(\sup_{\substack{\|\boldsymbol{\eta}\|_2=1 \\ \frac{b}{2}\leqslant\|\boldsymbol{\eta}\|_1\leqslant b}} \left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i\langle \boldsymbol{\eta}, \boldsymbol{Z}_i\rangle\right|\right) \\
&\leqslant 8Kb\mathbb{E}_{\boldsymbol{Z},\boldsymbol{\varepsilon}}\left(\sup_{\|\bar{\boldsymbol{\eta}}\|_1\leqslant 1}\left|\left\langle \bar{\boldsymbol{\eta}}, \frac{1}{N}\sum_{i=1}^N \varepsilon_i \boldsymbol{Z}_i\right\rangle\right|\right) \qquad \text{(where } \overline{\boldsymbol{\eta}} := \boldsymbol{\eta}/\|\boldsymbol{\eta}\|_1\text{)} \\
&= 8Kb\mathbb{E}\left(\left\|\frac{1}{N}\sum_{i=1}^N \varepsilon_i \boldsymbol{Z}_i\right\|_\infty\right) = 8Kb\mathcal{R}_N,
\end{aligned}$$

where the final step follows by applying Hölder's inequality. Then choosing $c_2 = 32K$ gives,

$$\begin{aligned}
\mathbb{P}\left(Y_n(b) > \frac{1}{2}\kappa_1 + \frac{1}{2}c_2\mathcal{R}_N b\right) &\leqslant \mathbb{P}\left(Y_n(b) - \mathbb{E}(Y_n(b)) > \frac{1}{2}\kappa_1 + 8Kb\mathcal{R}_N\right) \\
&\leqslant e^{-O_{\kappa_1,\kappa_2}(n)},
\end{aligned}$$

using (89) with $c_3 = 8K$. This proves Lemma 20 (2).

Final Details: Peeling Strategy

Recall from (86) that we have proved (85) holds for any fixed b such that $\frac{b}{2} \leqslant \|\boldsymbol{\eta}\|_1 \leqslant b$ and $\|\boldsymbol{\eta}\|_2 \leqslant 1$. The final step in the proof of Proposition 19 is to remove the restriction on the ℓ_1 norm of $\boldsymbol{\eta}$. For this, we apply a peeling strategy as in the proof of Wainwright (2019, Theorem 9.34). To this end, consider the set

$$\mathbb{S}_\ell := \left\{ \boldsymbol{\eta} \in \mathbb{R}^d : 2^{\ell-1} \leqslant \frac{\|\boldsymbol{\eta}\|_1}{\|\boldsymbol{\eta}\|_2} \leqslant 2^\ell \right\} \cap \left\{ \boldsymbol{\eta} \in \mathbb{R}^d : \|\boldsymbol{\eta}\|_2 \leqslant 1 \right\},$$

for $\ell = 1, \ldots, \lceil \log_2(\sqrt{d}) \rceil$. Then for $\boldsymbol{\eta} \in \mathbb{R}^d \cap \mathbb{S}_\ell$ by (86),

$$\frac{1}{N} \sum_{i=1}^N \langle \boldsymbol{\eta}, \boldsymbol{Z}_i \rangle^2 \psi(\langle \boldsymbol{\theta}, \boldsymbol{Z}_i \rangle + t \langle \boldsymbol{\eta}, \boldsymbol{Z}_i \rangle) \geqslant \nu \|\boldsymbol{\eta}\|_2^2 - c_0 \mathcal{R}_N \|\boldsymbol{\eta}\|_1^2,$$

with probability at least $1 - e^{-O_{\kappa_1,\kappa_2}(n)}$. Then by a union bound,

$$\frac{1}{N} \sum_{i=1}^N \langle \boldsymbol{\eta}, \boldsymbol{Z}_i \rangle^2 \psi(\langle \boldsymbol{\theta}, \boldsymbol{Z}_i \rangle + t \langle \boldsymbol{\eta}, \boldsymbol{Z}_i \rangle) \geqslant \nu \|\boldsymbol{\eta}\|_2^2 - c_0 \mathcal{R}_N \|\boldsymbol{\eta}\|_1^2,$$

for all $\boldsymbol{\eta} \in \mathbb{R}^d$ such that $\|\boldsymbol{\eta}\|_2 \leqslant 1$, with probability at least $1 - \lceil \log_2(d) \rceil e^{-O_{\kappa_1,\kappa_2}(n)} = 1 - o(1)$. This completes the proof of Proposition 19.

Appendix C. Proofs from Section 3

In this section, we prove the results stated in Section 3. We begin with the proof of Theorem 3 in Section C.1. Corollary 5 is proved in Section C.2.

C.1 Proof of Theorem 3

The proof of Theorem C.1 follows along the same lines as in Theorem 1, so to avoid repetition we sketch the steps and highlight the relevant modifications. As in the proof of Theorem C.1, the first step is to show the concentration of $\nabla L_N(\boldsymbol{\gamma})$. Towards this, following the proof of Lemma 13 shows that Q_r (as defined in (36)) is $O_\beta(\mathrm{poly}(d_{\max})/N)$-Lipschitz, for each $r \in [\ell]$. Therefore, by arguments as in Lemma 13,

$$\mathbb{P}\left(\left| \sum_{i=1}^N \phi_i(\boldsymbol{X}) \right| \geqslant t \right) \leqslant \mathbb{P}\left(\max_{r \in [\ell]} |Q_r(\boldsymbol{X})| \geqslant \frac{t\ell'}{\ell} \right) \lesssim e^{-O_{\beta,M}(Nt^2/\mathrm{poly}(d_{\max}))}.$$

since $\ell = O(\log N)$ and $\ell'/\ell = \Theta(1)$. In this case, following the notations in the proof of Lemma 8, we have $\ell/\ell' = \Theta(d_{\max})$. Similarly, for $s \in [d]$,

$$\mathbb{P}\left(\left| \sum_{i=1}^N \phi_{i,s}(\boldsymbol{X}) \right| \geqslant t \right) \lesssim e^{-O_{\beta,M}(Nt^2/\mathrm{poly}(d_{\max}))}.$$

Hence, choosing $\lambda := \delta \text{poly}(d_{\max})\sqrt{\log(d+1)/N}$ and $t := \lambda/2$ gives, for some constant C depending only on β and M,

$$\mathbb{P}\left(\left|\sum_{i=1}^{N}\phi_i(\boldsymbol{X})\right| \geqslant \frac{\lambda}{2}\right) \leqslant (d+1)^{-C\delta^2/4} \quad \text{and} \quad \mathbb{P}\left(\left|\sum_{i=1}^{N}\phi_{i,j}(\boldsymbol{X})\right| \geqslant \frac{\lambda}{2}\right) \leqslant (d+1)^{-C\delta^2/4},$$

for all $j \in [s]$. A final union bound over the $(d+1)$ coordinates now shows that

$$\mathbb{P}\left(\|\nabla L_N(\boldsymbol{\gamma})\|_\infty > \frac{\lambda}{2}\right) = o(1), \tag{90}$$

where $\lambda := \delta \text{poly}(d_{\max})\sqrt{\log(d+1)/N}$ as above and the $o(1)$-term goes to infinity as $d \to \infty$. This establishes the concentration of the gradient.

Next, we need to show the strong concavity of the Hessian. To this end, first note that since $|E(G_N)| = O(N)$ and the number of non-isolated vertices of G_N is $\Omega(N)$, there exist constants $L_1, L_2 > 0$, such that $|E(G_N)| \leqslant L_1 N$, and the number of non-isolated vertices of G_N is larger than $L_2 N$, for all N large enough. For $D \geqslant 1$ define,

$$V_N(D) := \{v \in V(G_N) : d_v \in [1, D]\}.$$

Note that

$$|V_N(D)| \geqslant N(1 - (2L_1/D)) - N(1 - L_2) = N(L_2 - (2L_1/D)),$$

since G_N has at least $N(1-(2L_1/D))$ vertices with degree not exceeding D, among which at most $N(1-L_2)$ are isolated. Hereafter, we choose $D := \lceil 4L_1/L_2 \rceil$, so that $|V_N(D)| \geqslant L_2 N/2$. Then, with notations as in (45) we have,

$$\begin{aligned}
&L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\
&= \frac{1}{2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \nabla^2 L_N(\underline{\boldsymbol{\gamma}})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\
&= \frac{1}{2N}\sum_{i=1}^{N} \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{U}_i \boldsymbol{U}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\cosh^2(\underline{\beta} m_i(\boldsymbol{X}) + \underline{\boldsymbol{\theta}}^\top \boldsymbol{Z}_i)} \\
&\geqslant \frac{1}{2N}\sum_{i \in V_N(D)} \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{U}_i \boldsymbol{U}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\cosh^2(\underline{\beta} m_i(\boldsymbol{X}) + \underline{\boldsymbol{\theta}}^\top \boldsymbol{Z}_i)} \\
&\geqslant \frac{1}{2N}\sum_{i \in V_N(D)} \frac{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{U}_i \boldsymbol{U}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})}{\cosh^2(|\beta| D + (D+M)\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 + sM\Theta)},
\end{aligned} \tag{91}$$

where the last step uses the bound in (47) and the bound $|\underline{\beta} m_i(\boldsymbol{X})| \leqslant |\underline{\beta}||m_i(\boldsymbol{X})| \leqslant |\underline{\beta}| D \leqslant |\beta| D + D\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1$. Next, using the bound $|V_N(D)| \geqslant L_2 N/2$ in (91) gives,

$$\begin{aligned}
&L_N(\hat{\boldsymbol{\gamma}}) - L_N(\boldsymbol{\gamma}) - \nabla L_N(\boldsymbol{\gamma})^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\
&\geqslant \frac{L_2}{4\cosh^2(|\beta| D + (D+M)\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 + sM\Theta)} \cdot \frac{1}{|V_N(D)|}\sum_{i \in V_N(D)} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \boldsymbol{U}_i \boldsymbol{U}_i^\top (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \\
&= \frac{L_2}{4\cosh^2(|\beta| D + (D+M)\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 + sM\Theta)} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^\top \tilde{\boldsymbol{G}} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}),
\end{aligned} \tag{92}$$

where

$$\tilde{\boldsymbol{G}} := \frac{1}{|V_N(D)|} \begin{pmatrix} \tilde{\boldsymbol{m}}^\top \tilde{\boldsymbol{m}} & \tilde{\boldsymbol{m}}^\top \tilde{\boldsymbol{Z}} \\ \tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{m}} & \tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}} \end{pmatrix},$$

with $\tilde{\boldsymbol{m}} := (m_i(\boldsymbol{X}))_{i \in V_N(D)}^\top$ and $\tilde{\boldsymbol{Z}} = (\boldsymbol{Z}_i)_{i \in V_N(D)}^\top$. The calculation in (92) implies that in order to establish the strong concavity of the Hessian in the setting of Theorem 3, we need to show $\lambda_{\min}(\tilde{\boldsymbol{G}}) \gtrsim 1$ with probability going to 1, where $\lambda_{\min}(\tilde{\boldsymbol{G}})$ denotes the minimum eigenvalue of $\tilde{\boldsymbol{G}}$. For this step, we follow the steps of Lemma 14 with the matrix $\boldsymbol{F}$ replaced by $\tilde{\boldsymbol{F}} := \boldsymbol{I} - \tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}})^{-1} \tilde{\boldsymbol{Z}}^\top$. Then, repeating the proof of (63) in Lemma 17, with $\boldsymbol{A}$ replaced by $\tilde{\boldsymbol{A}} := \boldsymbol{A}|_{V_N(D) \times [N]}$, we can find a set $J \subseteq [N]$ such that

$$\mathbb{E}\left(\|\tilde{\boldsymbol{F}} \tilde{\boldsymbol{m}}\|_2^2 \Big| \boldsymbol{X}_{J^c} \right) \gtrsim \Upsilon^2 \left(\|\tilde{\boldsymbol{A}}\|_F^2 - d \cdot \mathrm{polylog}(N) \right),$$

since $\|\tilde{\boldsymbol{A}}\|_2 \leqslant \|\boldsymbol{A}\|_\infty = O(\mathrm{polylog}(N))$. This implies,

$$\mathbb{E}\left(\|\tilde{\boldsymbol{F}} \tilde{\boldsymbol{m}}\|_2^2 \Big| \boldsymbol{X}_{J^c} \right) \gtrsim N, \tag{93}$$

since $\|\tilde{\boldsymbol{A}}\|_F^2 \geqslant |V_N(D)| \gtrsim N$ (because every vertex in $V_N(D)$ has degree at least 1) and $d = o(N)$. The final step is to establish that $\|\tilde{\boldsymbol{F}} \tilde{\boldsymbol{m}}\|$ concentrates around $\|\tilde{\boldsymbol{F}} \tilde{\boldsymbol{m}}\|_2^2 \Big| \boldsymbol{X}_{J^c}$, conditional on $\boldsymbol{X}_{J^c}$. This follows by repeating the proof of Lemma 18, which introduces an extra $1/\mathrm{polylog}(N)$ factor in each of the two exponential terms in the RHS of (64), since $\|\tilde{\boldsymbol{A}}\|_2 \leqslant \|\boldsymbol{A}\|_\infty = O(\mathrm{polylog}(N))$. This, combined with (93), shows

$$\mathbb{P}(\lambda_{\min}(\tilde{\boldsymbol{G}}) \geqslant C) \geqslant 1 - e^{-\Omega(N/\mathrm{polylog}(N))}, \tag{94}$$

for some constant $C > 0$. The proof of Theorem 3 can be now completed using (90) and (94), as in Theorem 1.

C.2 Proof of Corollary 5

To prove Corollary 5 we verify that the hypotheses of Theorem 3 are satisfied. Note that we can write

$$|E(G_N)| = \sum_{1 \leqslant u < v \leqslant N} B_{uv},$$

where $B_{uv} \sim \mathrm{Ber}(p_{uv})$, and $\{B_{uv}\}_{1 \leqslant u < v \leqslant N}$ are independent. This implies, $\mathbb{E}|E(G_N)| = O(N)$ and $\mathrm{Var}(|E(G_N)|) = O(N)$, since $\sup_{1 \leqslant i,j \leqslant N} p_{ij} = O(1/N)$ by (15). Hence, by Chebyshev's inequality,

$$|E(G_N)| \leqslant \mathbb{E}|E(G_N)| + N = O(N),$$

with probability $1 - o(1)$.

Next, we will show that $d_{\max} = \tilde{O}(1)$ holds with high probability. To this end, define $\eta_u = \sum_{v \neq u} p_{uv}$, for $1 \leqslant u \leqslant N$. Clearly, by assumption (15), $\max_{1 \leqslant u \leqslant N} \eta_u = O(1)$. Next, we establish that $\max_{1 \leqslant u \leqslant N} \eta_u = \Omega(1)$. Towards this, note that by assumption (16) there exists $\varepsilon \in (0,1)$ and $u \in V(G_N)$ such that,

$$\limsup_{N \to \infty} \sum_{v=1}^{N} \log(1 - p_{uv}) < \log \varepsilon.$$

Next, using the inequality $\log(1-x) \geqslant -x/(1-x)$, for all $x < 1$, and taking N large enough such that $\sup_{u,v} p_{uv} < 1/2$, gives

$$\limsup_{N\to\infty} \left(-2\sum_{v=1}^{N} p_{uv}\right) \leqslant \log \varepsilon \quad \Longrightarrow \quad \liminf_{N\to\infty} \eta_u \geqslant \frac{1}{2}\log\left(\frac{1}{\varepsilon}\right).$$

This shows that $\max_{1\leqslant u\leqslant N} \eta_u = \Omega(1)$. Therefore, by Proposition 1.11 in Benaych-Georges et al. (2019), $d_{\max} \leqslant O(\log N)$ with probability $1 - o(1)$.

Finally, we show that the number of non-isolated vertices of G_N is $\Omega(N)$ with high probability. To this end, for each $v \in V(G_N)$, define $Y_v := \mathbf{1}\{d_v = 0\}$. Then, $I_N := \sum_{v=1}^{N} Y_v$ is the total number of isolated vertices of G_N. Note that $Y_v \sim \text{Ber}(\prod_{u=1}^{N}(1-p_{uv}))$. Therefore, by (16),

$$\mathbb{E}(I_N) = \sum_{u=1}^{N}\prod_{v=1}^{N}(1-p_{uv}) \leqslant \alpha N,$$

for some $\alpha \in (0,1)$ and all large N enough. Next, note that for any two distinct vertices $u, v \in V(G_N)$,

$$\text{Cov}(Y_u, Y_v) = p_{uv}(1-p_{uv}) \prod_{w\notin\{u,v\}} [(1-p_{uw})(1-p_{vw})] \leqslant p_{uv} = O\left(\frac{1}{N}\right).$$

Similarly, it can be checked that $\text{Var}(Y_v) = O(1)$, for $1 \leqslant v \leqslant N$. This implies, $\text{Var}(I_N) = O(N)$. Hence, by Chebyshev's inequality,

$$\mathbb{P}\left(I_N \geqslant \left(\frac{1+\alpha}{2}\right)N\right) \leqslant \mathbb{P}\left(I_N \geqslant \mathbb{E}(I_N) + \left(\frac{1-\alpha}{2}\right)N\right) \leqslant \frac{\text{Var}(I_N)}{\left(\frac{1-\alpha}{2}\right)^2 N^2} = O\left(\frac{1}{N}\right).$$

This shows that the number of non-isolated vertices of G_N is at least $(1-\alpha)N/2$ with probability $1 - o(1)$. This completes the verification of the hypotheses of Theorem 3 and hence, Corollary 5 follows from Theorem 3.

Appendix D. Proofs of Technical Lemmas

In this section we collect the proofs of various technical lemmas. The section is organized as follows: In Appendix D.1 we prove Lemma 13. The proof of Lemma 11 is given in Appendix D.2. In Appendix D.3 we prove a variance lower bound for linear functions.

D.1 Proof of Lemma 13

To begin with, recall from (36) that

$$Q_r(\boldsymbol{X}) = -\frac{1}{N}\sum_{i\in I_r} m_i(\boldsymbol{X})\left[X_i - \tanh(\beta m_i(\boldsymbol{X}) + \boldsymbol{\theta}^\top \boldsymbol{Z}_i)\right].$$

Hence, for any two $\boldsymbol{X}, \boldsymbol{X}' \in \{-1,1\}^N$,

$$|Q_r(\boldsymbol{X}) - Q_r(\boldsymbol{X}')| = T_1 + T_2, \tag{95}$$

where

$$T_1 = \frac{1}{N}\left|\sum_{i\in I_r}\left\{m_i(\boldsymbol{X})X_i - m_i(\boldsymbol{X}')X_i'\right\}\right|$$
$$T_2 = \frac{1}{N}\left|\sum_{i\in I_r}\left\{m_i(\boldsymbol{X})\tanh(\beta m_i(\boldsymbol{X}) + \boldsymbol{\theta}^\top \boldsymbol{Z}_i) - m_i(\boldsymbol{X}')\tanh(\beta m_i(\boldsymbol{X}') + \boldsymbol{\theta}^\top \boldsymbol{Z}_i)\right\}\right|. \quad (96)$$

Now, assume that $\boldsymbol{X}$ and $\boldsymbol{X}'$ differ only in the k-th coordinate, for some $k \in [N]$. Then

$$\begin{aligned} T_1 &= \left|\sum_{i\in I_r}\sum_{j=1}^N a_{ij}(X_iX_j - X_i'X_j')\right| \\ &\leqslant \left|\sum_{i\in I_r} a_{ik}(X_iX_k - X_i'X_k')\right| + \left|\sum_{j=1}^N a_{kj}(X_kX_j - X_k'X_j')\right| \\ &\leqslant 2\sum_{i\in I_r}|a_{ik}| + 2\sum_{j=1}^N |a_{kj}| \\ &\leqslant 4\left|\sum_{i=1}^N a_{ik}\right| \leqslant 4\|\boldsymbol{A}\|_1 \leqslant 4 \quad \text{(by (23))}. \end{aligned} \quad (97)$$

Next, we proceed to bound T_2. Towards this note that

$$T_2 \leqslant T_{21} + T_{22}, \quad (98)$$

where

$$T_{21} := \left|\sum_{i\in I_r} m_i(\boldsymbol{X})\left\{\tanh(\beta m_i(\boldsymbol{X}) + \boldsymbol{\theta}^\top \boldsymbol{Z}_i) - \tanh(\beta m_i(\boldsymbol{X}') + \boldsymbol{\theta}^\top \boldsymbol{Z}_i)\right\}\right|$$
$$T_{22} := \left|\sum_{i\in I_r} a_{ik}(X_k - X_k')\tanh(\beta m_i(\boldsymbol{X}') + \boldsymbol{\theta}^\top \boldsymbol{Z}_i)\right|,$$

since $m_i(\boldsymbol{X}) - m_i(\boldsymbol{X}') = a_{ik}(X_k - X_k')$. Hence, using $|\tanh x - \tanh y| \leqslant |x - y|$ gives,

$$T_{21} \leqslant |\beta|\sum_{i\in I_r}|m_i(\boldsymbol{X})|\left|m_i(\boldsymbol{X}) - m_i(\boldsymbol{X}')\right| \leqslant 2|\beta|\sum_{i\in I_r}|m_i(\boldsymbol{X})||a_{ik}| \leqslant 2|\beta|\|\boldsymbol{A}\|_\infty\|\boldsymbol{A}\|_1 \leqslant 2|\beta|. \quad (99)$$

and using $\tanh x \leqslant 1$ gives,

$$T_{22} \leqslant 2\sum_{i\in I_r}|a_{ir}| \leqslant 2\|\boldsymbol{A}\|_1 \leqslant 2. \quad (100)$$

Combining (95), (96), (97), (98), (99), (100) the result in Lemma 13 (1) follows.

Next, we prove (2). To begin with, recall from (38) that

$$Q_{r,s}(\boldsymbol{X}) = -\frac{1}{N}\sum_{i\in I_r} \boldsymbol{Z}_{i,s}\left[X_i - \tanh(\beta m_i(\boldsymbol{X}) + \boldsymbol{\theta}^\top \boldsymbol{Z}_i)\right].$$

Hence, for any two $\boldsymbol{X}, \boldsymbol{X}' \in \{-1,1\}^N$ differing in the k-th coordinate only,

$$\begin{aligned}
|Q_{r,s}(\boldsymbol{X}) - Q_{r,s}(\boldsymbol{X}')| &\leqslant \frac{1}{N}\left|\sum_{i\in I_r} Z_{i,s}(X_i - X_i')\right| \\
&\quad + \frac{1}{N}\left|\sum_{i\in I_r} Z_{i,s}\left[\tanh(\beta m_i(\boldsymbol{X}) + \boldsymbol{\theta}^\top \boldsymbol{Z}_i) - \tanh(\beta m_i(\boldsymbol{X}') + \boldsymbol{\theta}^\top \boldsymbol{Z}_i)\right]\right| \\
&\leqslant \frac{2|Z_{k,s}|}{N} + \frac{|\beta|}{N}\left|\sum_{i\in I_r} Z_{i,s}(m_i(\boldsymbol{X}) - m_i(\boldsymbol{X}'))\right| \\
&\leqslant \frac{2|Z_{k,s}|}{N} + \frac{2|\beta|}{N}\sum_{i=1}^N |Z_{i,s}a_{ik}|,
\end{aligned}$$

as desired.

D.2 Proof of the Lemma 11

Suppose that $\boldsymbol{X}$ comes from the model:

$$\mathbb{P}_{\beta,\boldsymbol{h}}(\boldsymbol{X}) \;\propto\; \exp\left(\sum_{i=1}^N h_i X_i + \boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X}\right), \tag{101}$$

which is assumed to be (R,Υ)-Ising.We will apply Lemma 17 in Dagan et al. (2021) on the matrix

$$\boldsymbol{D} := ((d_{ij}))_{1\leqslant i,j\leqslant N} := \boldsymbol{A}/R\,,$$

where $\boldsymbol{A}$ is the interaction matrix corresponding to the distribution of $\boldsymbol{X}$. Note that $\boldsymbol{D}$ satisfies the hypotheses of Lemma 17 in Dagan et al. (2021). For $\eta \in (0,R)$, define $\eta' := \eta/R$. This ensures that $\eta' \in (0,1)$. By Lemma 17 in Dagan et al. (2021), there exist subsets $I_1,\ldots,I_\ell \subseteq [N]$ with $\ell \lesssim R^2\log N/\eta^2$, such that for all $1\leqslant i\leqslant N$, $|\{j\in\ell : i\in I_j\}| = \lceil \eta\ell/8R\rceil$, and for all $j\in\ell$,

$$\|\boldsymbol{D}|_{I_j\times I_j}\|_\infty \leqslant \eta' \implies \|\boldsymbol{A}|_{I_j\times I_j}\|_\infty \leqslant \eta, \tag{102}$$

where for a matrix $\boldsymbol{M} = ((m_{ij})) \in \mathbb{R}^{s\times t}$ and for sets $S\subseteq\{1,\ldots,s\}, T\subseteq\{1,\ldots,t\}$, we define $\boldsymbol{M}|_{S\times T} := ((m_{ij}))_{i\in S, j\in T} \in \mathbb{R}^{|S|\times|T|}$.

Now, for $j\in[\ell]$,

$$\begin{aligned}
&\frac{\mathbb{P}(\boldsymbol{X}_{I_j} = \boldsymbol{y}|\boldsymbol{X}_{I_j^c} = \boldsymbol{x}_{-I_j})}{\mathbb{P}(\boldsymbol{X}_{I_j} = \boldsymbol{y}'|\boldsymbol{X}_{I_j^c} = \boldsymbol{x}_{-I_j})} \\
&= \frac{\exp\left(\boldsymbol{y}_{I_j}^\top \boldsymbol{D}|_{I_j\times I_j}\boldsymbol{y}_{I_j} + \sum_{u\in I_j, v\notin I_j} d_{uv}y_u x_v + \sum_{i\in I_j} h_i y_i\right)}{\exp\left(\boldsymbol{y}_{I_j}^\top \boldsymbol{D}|_{I_j\times I_j}\boldsymbol{y}_{I_j} + \sum_{u\in I_j, v\notin I_j} d_{uv}y'_u x_v + \sum_{i\in I_j} h_i y'_i\right)}.
\end{aligned} \tag{103}$$

Observe that the RHS of (103) is the probability mass function of an Ising model μ on $\{-1,1\}^{|I_j|}$ with interaction matrix $\boldsymbol{D}|_{I_j\times I_j}$ and external magnetic field term at site i given by

$$h_i' = \sum_{v\notin I_j} D_{iv}x_v + h_i.$$

Recall from (102) that $\|\boldsymbol{A}|_{I_j\times I_j}\|_\infty \leqslant \eta$. The next step is to show that if $\boldsymbol{Y}\sim\mu$, then

$$\min_{1\leqslant u\leqslant |I_j|} \mathrm{Var}(Y_u|\boldsymbol{Y}_{-u}) \geqslant \Upsilon\ .$$

Towards this, note that for any $1\leqslant u\leqslant |I_j|$,

$$\mathbb{P}(Y_u=1|\boldsymbol{Y}_{-u}=\boldsymbol{y}_{-u}) = \mathbb{P}(X_{I_j^u}=1|\boldsymbol{X}_{I_j\setminus\{I_j^u\}}=\boldsymbol{y}_{-u}, \boldsymbol{X}_{-I_j}=\boldsymbol{x}_{-I_j})$$

where I_j^u is the u-th smallest element of I_j. Hence,

$$\mathrm{Var}(Y_u|\boldsymbol{Y}_{-u}=\boldsymbol{y}_{-u}) = \mathrm{Var}(X_{I_j^u}|\boldsymbol{X}_{I_j\setminus\{I_j^u\}}=y_{-u}, \boldsymbol{X}_{-I_j}=\boldsymbol{x}_{-I_j}) \geqslant \Upsilon,$$

since (101) is a (R,Υ)-Ising model. This implies that $X_{I_j}|X_{I_j^c}$ is (η,Υ)-Ising model, for $j\in[\ell]$.

To prove (33), let us pick $j\in[\ell]$ uniformly at random, and note that for any vector $\boldsymbol{a}$,

$$\mathbb{E}\left(\sum_{i\in I_j} a_i\right) = \sum_{i=1}^N a_i\mathbb{P}(I_j\ni i) = \sum_{i=1}^N a_i\frac{\lceil\eta\ell/8R\rceil}{\ell} \geqslant \frac{\eta}{8R}\sum_{i=1}^N a_i.$$

This means that there is a fixed sample point j, such that

$$\sum_{i\in I_j} a_i \geqslant \frac{\eta}{8R}\sum_{i=1}^N a_i.$$

This completes the proof of Lemma 11.

D.3 Variance Lower Bound for Linear Functions

In this section we derive a variance lower bound for linear functions in (R,Υ)-Ising models. This follows the proof of Lemma 10 in Dagan et al. (2021) adapted to our setting. We begin with the following definition: For two probability measures μ and ν, the ℓ_1-Wasserstein distance is defined as:

$$W_1(\mu,\nu) := \min_{\pi\in\mathscr{C}_{\mu,\nu}} \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\pi}\|\boldsymbol{U}-\boldsymbol{V}\|_1,$$

where $\mathscr{C}_{\mu,\nu}$ denotes the set of all couplings of the probability measures μ and ν.

Lemma 22 *Let $\boldsymbol{X}\in\{-1,1\}^N$ be a sample from an (R,Υ)- Ising model for some $R<1/8$. Then, for each $i\in[N]$,*

$$W_1\left(\mathbb{P}_{\boldsymbol{X}_{-i}|X_i=1},\ \mathbb{P}_{\boldsymbol{X}_{-i}|X_i=-1}\right) \leqslant \frac{16R}{1-8R}.$$

Proof It follows from Lemma 4.9 in Dagan et al. (2019), that there exists a coupling π of the conditional measures $\mathbb{P}_{\boldsymbol{X}_{-i}|X_i=1}$ and $\mathbb{P}_{\boldsymbol{X}_{-i}|X_i=-1}$, such that:

$$\mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\pi}\left[d_H(\boldsymbol{U},\boldsymbol{V})\right] \leqslant \frac{\alpha}{1-\alpha}, \tag{104}$$

where d_H denotes the Hamming distance and α the Dobrushin coefficient. By Lemma 4.4 in Chatterjee (2016), we know that Dobrushin's interdependence matrix is given by $8\boldsymbol{D}$. Hence, it follows from (104) that (Dobrushin's coefficient in Theorem 2.3 in Dagan et al. (2019) is given by $\alpha = 8\|\boldsymbol{D}\|_2$, see Theorem 4.3 in Chatterjee (2016)),

$$W_1\left(\mathbb{P}_{\boldsymbol{X}_{-i}|X_i=1},\ \mathbb{P}_{\boldsymbol{X}_{-i}|X_i=-1}\right) \leqslant 2\ \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\pi}\left[d_H(\boldsymbol{U},\boldsymbol{V})\right] \leqslant \frac{16\|\boldsymbol{D}\|_2}{1-8\|\boldsymbol{D}\|_2} \leqslant \frac{16R}{1-8R}.$$

which completes the proof of the lemma. ■

Using the above lemma, we now prove the desired variance lower bound:

Lemma 23 *Let $\boldsymbol{X}$ be a sample from an (R,Υ)-Ising model for some $R>0$. Then for any vector $\boldsymbol{a}\in\mathbb{R}^N$,*

$$\mathrm{Var}(\boldsymbol{a}^\top\boldsymbol{X}) \gtrsim \frac{\|\boldsymbol{a}\|_2^2\Upsilon^2}{R}.$$

Proof First, consider the case $R \leqslant \Upsilon/32 \leqslant 1/32$. In this case, $1-8R \geqslant 3/4$, so

$$\frac{8R}{1-8R} \leqslant \frac{\Upsilon}{3} \tag{105}$$

Now, it follows from Lemma 22 that for every i,

$$\begin{aligned}\sum_{j\in[N]\setminus\{i\}} \left|\mathbb{E}\left(X_j|X_i=1\right)-\mathbb{E}\left(X_j|X_i=-1\right)\right| &\leqslant \sum_{j\in[N]\setminus\{i\}} W_1\left(\mathbb{P}_{X_j|X_i=1},\mathbb{P}_{X_j|X_i=-1}\right)\Big)\\ &\leqslant W_1\left(\mathbb{P}_{\boldsymbol{X}_{-i}|X_i=1},\ \mathbb{P}_{\boldsymbol{X}_{-i}|X_i=-1}\right)\\ &\leqslant \frac{16R}{1-8R}.\end{aligned} \tag{106}$$

Next, note that:

$$\begin{aligned}\mathrm{Cov}(X_i,X_j) &= \mathbb{E}[(X_i-\mathbb{E}X_i)X_j]\\ &= \mathbb{E}\left[(X_i-\mathbb{E}X_i)\mathbb{E}(X_j|X_i)\right]\\ &= \mathbb{P}(X_i=1)(1-\mathbb{E}X_i)\mathbb{E}(X_j|X_i=1)-\mathbb{P}(X_i=-1)(1+\mathbb{E}X_i)\mathbb{E}(X_j|X_i=-1)\\ &= \frac{1+\mathbb{E}X_i}{2}(1-\mathbb{E}X_i)\mathbb{E}(X_j|X_i=1)-\frac{1-\mathbb{E}X_i}{2}(1+\mathbb{E}X_i)\mathbb{E}(X_j|X_i=-1)\\ &= \frac{1}{2}[1-(\mathbb{E}X_i)^2]\left[\mathbb{E}(X_j|X_i=1)-\mathbb{E}(X_j|X_i=-1)\right]\\ &\leqslant \frac{1}{2}\left[\mathbb{E}(X_j|X_i=1)-\mathbb{E}(X_j|X_i=-1)\right]\end{aligned} \tag{107}$$

Combining (105), (106) and (107), we have for every i,

$$\sum_{j\in[N]\setminus\{i\}} |\mathrm{Cov}(X_i, X_j)| \leqslant \frac{8R}{1-8R} \leqslant \frac{\Upsilon}{3}. \tag{108}$$

Hence, we have by (108),

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{X}) &\geqslant \sum_{i=1}^N a_i^2 \mathrm{Var}(X_i) - \sum_{i\neq j} |a_i a_j \mathrm{Cov}(X_i, X_j)| \\
&\geqslant \sum_{i=1}^N a_i^2 \mathrm{Var}(X_i) - \sum_{i\neq j} \frac{(a_i^2 + a_j^2)\,|\mathrm{Cov}(X_i, X_j)|}{2} \\
&= \sum_{i=1}^N a_i^2 \left[\mathrm{Var}(X_i) - \sum_{j\in[N]\setminus\{i\}} |\mathrm{Cov}(X_i, X_j)|\right] \\
&\geqslant \sum_{i=1}^N a_i^2 \left(\Upsilon - \frac{\Upsilon}{3}\right) \\
&= \frac{2\Upsilon}{3}\|\boldsymbol{a}\|_2^2 \geqslant \frac{\|\boldsymbol{a}\|_2^2 \Upsilon^2}{R} \cdot \frac{2R}{3} \gtrsim \frac{\|\boldsymbol{a}\|_2^2 \Upsilon^2}{R}.
\end{aligned} \tag{109}$$

Now consider the case $R > \Upsilon/32$. By Lemma 11, we choose a subset I of $[N]$ such that conditioned on $\boldsymbol{X}_{I^c}$, $\boldsymbol{X}_I$ is a $(\Upsilon/32, \Upsilon)$- Ising model, and

$$\|\boldsymbol{a}_I\|_2^2 \geqslant \frac{\Upsilon}{256R}\|\boldsymbol{a}\|_2^2. \tag{110}$$

Hence, we have from (109) and (110),

$$\mathrm{Var}\left(\boldsymbol{a}^\top \boldsymbol{X} | \boldsymbol{X}_{I^c}\right) = \mathrm{Var}\left(\boldsymbol{a}_I^\top \boldsymbol{X}_I | \boldsymbol{X}_{I^c}\right) \geqslant \frac{2\Upsilon\|\boldsymbol{a}_I\|_2^2}{3} \geqslant \frac{\Upsilon^2\|\boldsymbol{a}\|_2^2}{384R}. \tag{111}$$

Lemma 23 now follows from (111) on observing that $\mathrm{Var}(\boldsymbol{a}^\top \boldsymbol{X}) \geqslant \mathbb{E}\left[\mathrm{Var}\left(\boldsymbol{a}^\top \boldsymbol{X} | \boldsymbol{X}_{I^c}\right)\right]$. ■

D.4 Lipschitz Condition for the Gradient of L_N

In this section we show that ∇L_N, the gradient of the pseudo-likelihood loss function L_N defined in (5), is Lipschitz.

Lemma 24 *Suppose the design matrix $\boldsymbol{Z} := (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)^\top$ satisfies $\lambda_{\max}\left(\frac{1}{N}\boldsymbol{Z}^\top \boldsymbol{Z}\right) = O(1)$. Then there exists a constant $L > 0$ such that for any two $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^{d+1}$,*

$$||\nabla L_N(\boldsymbol{\gamma}_1) - \nabla L_N(\boldsymbol{\gamma}_2)||_2 \leqslant L||\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2||_2.$$

Proof For any two $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^{d+1}$, there exists $\boldsymbol{\gamma}^* \in \mathbb{R}^{d+1}$ such that

$$\nabla L_N(\boldsymbol{\gamma}_1) - \nabla L_N(\boldsymbol{\gamma}_2) = (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2)^\top \nabla^2 L_N(\boldsymbol{\gamma}^*).$$

Therefore,

$$||\nabla L_N(\boldsymbol{\gamma}_1) - \nabla L_N(\boldsymbol{\gamma}_2)||_2 \leqslant \sup_{\boldsymbol{\gamma}^* \in \mathbb{R}^{d+1}} \lambda_{\max}(\nabla^2 L_N(\boldsymbol{\gamma}^*))||\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2||_2, \tag{112}$$

where $\lambda_{\max}(\nabla^2 L_N(\boldsymbol{\gamma}^*))$ is the largest eigenvalue of the Hessian matrix $\nabla^2 L_N(\boldsymbol{\gamma}^*)$. Recall from (45) that

$$\nabla^2 L_N(\boldsymbol{\gamma}^*) = \frac{1}{N}\sum_{i=1}^{N} \frac{\boldsymbol{U}_i \boldsymbol{U}_i^\top}{\cosh^2(\beta^* m_i(\boldsymbol{X}) + (\boldsymbol{\theta}^*)^\top \boldsymbol{Z}_i)},$$

where $\boldsymbol{\gamma}^* = (\beta^*, (\boldsymbol{\theta}^*)^\top)^\top$ and $\boldsymbol{U}_i := (m_i(\boldsymbol{X}), \boldsymbol{Z}_i^\top)^\top$, for $1 \leqslant i \leqslant N$. Using $\operatorname{sech}^2(x) \leqslant 1$ it follows that

$$\sup_{\boldsymbol{\gamma}^* \in \mathbb{R}^{d+1}} \lambda_{\max}(\nabla^2 L_N(\boldsymbol{\gamma}^*)) \leqslant \lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{U}_i \boldsymbol{U}_i^\top\right), \tag{113}$$

Now, suppose $\boldsymbol{w} = (u, \boldsymbol{v}^\top)^\top \in \mathbb{R}^{d+1}$ be such that $||\boldsymbol{w}||_2 = 1$. Then

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}^\top \boldsymbol{U}_i \boldsymbol{U}_i^\top \boldsymbol{w} = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{w}^\top \boldsymbol{U}_i)^2 &= \frac{1}{N}\sum_{i=1}^{N} (u m_i(\boldsymbol{X}) + \boldsymbol{v}^\top \boldsymbol{Z}_i)^2 \\
&\lesssim \frac{1}{N}\sum_{i=1}^{N} \left\{u^2 (m_i(\boldsymbol{X}))^2 + (\boldsymbol{v}^\top \boldsymbol{Z}_i)^2\right\} \\
&\leqslant \frac{1}{N}\sum_{i=1}^{N} m_i(\boldsymbol{X})^2 + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v}^\top \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \boldsymbol{v}.
\end{aligned}$$

Note that, since $|m_i(\boldsymbol{X})| \leqslant ||\boldsymbol{A}||_\infty \leqslant 1$ (by (23)), for $1 \leqslant i \leqslant N$, we have $\frac{1}{N}\sum_{i=1}^{N} m_i(\boldsymbol{X})^2 \leqslant 1$. Moreover, $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v}^\top \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \boldsymbol{v} = \frac{1}{N}\boldsymbol{v}^\top \boldsymbol{Z}^\top \boldsymbol{Z} \boldsymbol{v} \leqslant \lambda_{\max}(\frac{1}{N}\boldsymbol{Z}^\top \boldsymbol{Z}) = O(1)$. This implies

$$\lambda_{\max}(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{U}_i \boldsymbol{U}_i^\top) = O(1),$$

hence by (112) and (113), there exists a constant $L > 0$ such that

$$||\nabla L_N(\boldsymbol{\gamma}_1) - \nabla L_N(\boldsymbol{\gamma}_2)||_2 \leqslant L||\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2||_2.$$

This completes the proof of Lemma 24. ■