Carnegie Mellon University
Research Showcase @ CMU

Computer Science Department | School of Computer Science

4-19-2008

High-Dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression

Pradeep Ravikumar
University of California - Berkeley

Martin J. Wainwright
University of California - Berkeley

John Lafferty
Carnegie Mellon University, lafferty@cs.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/compsci

Published In

.

This Technical Report is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Computer Science Department by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

High-Dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression

Pradeep Ravikumar[†]
pradeepr@stat.berkeley.edu

Martin J. Wainwright[*,†]
wainwrig@stat.berkeley.edu

Department of Statistics[†], and Department of EECS[*]
UC Berkeley, Berkeley, CA 94720

John D. Lafferty
lafferty@cs.cmu.edu

Departments of Computer Science and Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213

Technical Report, Department of Statistics, UC Berkekley
April 19, 2008

Abstract

We consider the problem of estimating the graph structure associated with a discrete Markov random field. We describe a method based on ℓ_1-regularized logistic regression, in which the neighborhood of any given node is estimated by performing logistic regression subject to an ℓ_1-constraint. Our framework applies to the high-dimensional setting, in which both the number of nodes p and maximum neighborhood sizes d are allowed to grow as a function of the number of observations n. Our main results provide sufficient conditions on the triple (n, p, d) for the method to succeed in consistently estimating the neighborhood of every node in the graph simultaneously. Under certain assumptions on the population Fisher information matrix, we prove that consistent neighborhood selection can be obtained for sample sizes $n = \Omega(d^3 \log p)$, with the error decaying as $\mathcal{O}(\exp(-Cn/d^3))$ for some constant C. If these same assumptions are imposed directly on the sample matrices, we show that $n = \Omega(d^2 \log p)$ samples are sufficient.

Keywords: Graphical models, Markov random fields, structure learning, ℓ_1-regularization, model selection, convex risk minimization, high-dimensional asymptotics, concentration.

1 Introduction

Undirected graphical models, also known as Markov random fields (MRFs), are used in a variety of domains, including artificial intelligence, natural language processing, image analysis, statistical physics, and spatial statistics, among others. A Markov random field (MRF) is specified by an undirected graph $G = (V, E)$, with vertex set $V = \{1, 2, \ldots, p\}$ and edge set $E \subset V \times V$. The structure of this graph encodes certain conditional independence assumptions among subsets of the p-dimensional discrete random variable $X = (X_1, X_2, \ldots, X_p)$, where variable X_i is associated with vertex $i \in V$. A fundamental problem is the *graphical*

model selection problem: given a set of n samples $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ from a Markov random field, estimate the structure of the underlying graph. The *sample complexity* of such an estimator is the minimal number of samples n, as a function of the graph size p and possibly other parameters such as the maximum node degree d, required for the probability of correct identification of the graph to converge to one. Another important property of any model selection procedure is its *computational complexity*.

Due to both its importance and difficulty, structure learning in random fields has attracted considerable attention. The absence of an edge in a graphical model encodes a conditional independence assumption. Constraint-based approaches (Spirtes et al., 2000) estimate these conditional independencies from the data using hypothesis testing, and then determine a graph that most closely represents those independencies. Each graph represents a model class of graphical models; learning a graph then is a model class selection problem. Score-based approaches combine a metric for the complexity of the graph, with a goodness of fit measure of the graph to the data (for instance, log-likelihood of the maximum likelihood parameters given the graph), to obtain a *score* for each graph. The score is then used together with a search procedure that generates candidate graph structures to be scored. The number of graph structures grows super-exponentially, however, and Chickering (1995) shows that this problem is in general NP-hard.

A complication for undirected graphical models is that typical score metrics involve the normalization constant (also called the partition function) associated with the Markov random field, which is intractable (#P) to compute for general undirected models. The space of candidate structures in scoring based approaches is thus typically restricted to either directed models—Bayesian networks—or simple undirected graph classes such as trees (Chow and Liu, 1968), polytrees (Dasgupta, 1999) and hypertrees (Srebro, 2003). Abbeel et al. (2006) propose a method for learning factor graphs based on local conditional entropies and thresholding, and analyze its behavior in terms of Kullback-Leibler divergence between the fitted and true models. They obtain a sample complexity that grows logarithmically in the graph size p, but the computational complexity grows at least as quickly as $\mathcal{O}(p^{d+1})$, where d is the maximum neighborhood size in the graphical model. This order of complexity arises from the fact that for each node, there are $\binom{p}{d} = \mathcal{O}(p^d)$ possible neighborhoods of size d for a graph with p vertices. Csiszár and Talata (2006) show consistency of a method that uses pseudo-likelihood and a modification of the BIC criterion, but this also involves a prohibitively expensive search.

In work subsequent to the initial conference version of this work (Wainwright et al., 2007), other researchers have also studied the problem of model selection in discrete Markov random fields. For the special case of bounded degree models, Bresler et al. (2008) describe a simple search-based method, and prove under relatively mild assumptions that it can recover the graph structure with $\Theta(d \log p)$ samples. However, in the absence of additional restrictions, the computational complexity of the method is $\mathcal{O}(p^{d+1})$. Santhanam and Wainwright (2008) analyze the information-theoretic limits of graphical model selection, providing both upper and lower bounds on various model selection procedures, but these methods also have prohibitive computational cost.

The main contribution of this paper is to present and analyze the computational and sample complexity of a simple method for graphical model selection. Our analysis is high-dimensional in nature, meaning that both the model dimension p as well as the maximum

neighborhood size d may tend to infinity as a function of the size n. Our main result shows that under mild assumptions, consistent neighborhood selection is possible with sample complexity $n = \Omega(d^3 \log p)$ and computational complexity $\mathcal{O}(\max\{n, p\}p^3)$, when applied to any graph with p vertices and maximum degree d. The basic approach is straightforward: it involves performing ℓ_1-regularized logistic regression of each variable on the remaining variables, and then using the sparsity pattern of the regression vector to infer the underlying neighborhood structure.

The technique of ℓ_1 regularization for estimation of sparse models or signals has a long history in many fields; for instance, see Tropp (2006) for one survey. A surge of recent work has shown that ℓ_1-regularization can lead to practical algorithms with strong theoretical guarantees (e.g., Candes and Tao (2006), Donoho and Elad (2003), Meinshausen and Bühlmann (2006), Ng (2004), Tropp (2006), Wainwright (2006), Zhao and Yu (2007)). Despite the well-known computational intractability of discrete MRFs, our method is computationally efficient; it involves neither computing the normalization constant (or partition function) associated with the Markov random field, nor a combinatorial search through the space of graph structures. Rather, it requires only the solution of standard convex programs, with an overall computational complexity of order $\mathcal{O}(\max\{p, n\}\, p^3)$ (Koh et al., 2007), and is thus well-suited to high dimensional problems. Conceptually, like the work of Meinshausen and Bühlmann (2006) on covariance selection in Gaussian graphical models, our approach can be understood as using a type of pseudo-likelihood, based on the local conditional likelihood at each node. In contrast to the Gaussian case, where the exact maximum likelihood estimate can be computed exactly in polynomial time, this use of a surrogate loss function is essential for discrete Markov random fields, given the intractability of computing the exact likelihood.

The remainder of this paper is organized as follows. We begin in Section 2 with background on discrete graphical models, the model selection problem, and logistic regression. In Section 3, we state our main result, develop some of its consequences, and provide a high-level outline of the proof. Section 4 is devoted to proving a result under stronger assumptions on the sample Fisher information matrix itself, whereas Section 5 provides concentration results linking the population matrices to the sample versions. In Section 6, we provide some experimental results to illustrate the practical performance of our method, and the close agreement between theory and practice, and we conclude in Section 7.

Notation: For the convenience of the reader, we summarize here notation to be used throughout the paper. We use the following standard notation for asymptotics: we write $f(n) = \mathcal{O}(g(n))$ if $f(n) \leq Kg(n)$ for some constant $K < \infty$, and $f(n) = \Omega(g(n))$ if $f(n) \geq K'g(n)$ for some constant $K' > 0$. The notation $f(n) = \Theta(g(n))$ means that $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$. Given a vector $x \in \mathbb{R}^d$ and parameter $q \in [1, \infty]$, we use $\|x\|_q$ to denote the usual ℓ_q norm. Given a matrix $X \in \mathbb{R}^{a \times b}$ and parameter $q \in [1, \infty]$, we use $|\!|\!|X|\!|\!|_q$ to denote the induced matrix-operator norm (viewed as a mapping from $\ell_q^b \to \ell_q^a$); see Horn and Johnson (1985). Two examples of particular importance in this paper are the spectral norm $|\!|\!|X|\!|\!|_2$, corresponding to the maximal singular value of X, and the ℓ_∞ matrix norm, given by $|\!|\!|X|\!|\!|_\infty = \max_{j=1,\ldots,a} \sum_{k=1}^b |X_{jk}|$. We make use of the bound $|\!|\!|X|\!|\!|_\infty \leq \sqrt{a}|\!|\!|X|\!|\!|_2$, for any square matrix $X \in \mathbb{R}^{a \times a}$.

2 Background and problem formulation

We begin by providing some background on Markov random fields, defining the problem of graphical model selection, and describing our method based on neighborhood logistic regression.

2.1 Markov random fields

Given an undirected graph G with vertex set $V = \{1, \ldots, p\}$ and edge set E, a Markov random field (MRF) consists of random vector $X = (X_1, X_2, \ldots, X_p)$, where the random variable X_s is associated with vertex $s \in V$. The random vector $X \in \mathcal{X}^p$ is said to be pairwise Markov with respect to the graph if its probability distribution factorizes as $\mathbb{P}(x) \propto \exp\left\{\sum_{(s,t)\in E} \phi_{st}(x_s, x_t)\right\}$, where each ϕ_{st} is a mapping from pairs $(x_s, x_t) \in \mathcal{X}_s \times \mathcal{X}_t$ to the real line. An important special case is the Ising model, in which $X_s \in \{-1, 1\}$ for each vertex $s \in V$, and $\phi_{st}(x_s, x_t) = \theta_{st} x_s x_t$ for some parameter $\theta^*_{st} \in \mathbb{R}$, so that the distribution takes the form

$$\mathbb{P}_{\theta^*}(x) = \frac{1}{Z(\theta^*)} \exp\left\{\sum_{(s,t)\in E} \theta^*_{st} x_s x_t\right\}. \tag{1}$$

The partition function $Z(\theta^*)$ ensures that the distribution sums to one. The Ising model has proven useful in many domains, including statistical physics, where it describes the behavior of gases or magnets, in computer vision for image segmentation, and in social network analysis.

2.2 Graphical model selection

Suppose that we are given collection $\{x^{(i)}\} = \{x^{(1)}, \ldots, x^{(n)}\}$ of n samples, where each p-dimensional vector $x^{(i)}$ is drawn in an i.i.d. manner from a distribution $\mathbb{P}_{\theta^*}$ of the form (1). It is convenient to view the parameter vector θ^* as a $\binom{p}{2}$-dimensional vector, indexed by pairs of distinct vertices, but non-zero *if and only if* the vertex pair (s, t) belongs to the unknown edge set E of the underlying graph G. The goal of *graphical model selection* is to infer the edge set E of the graphical model defining the probability distribution that generates the samples. In this paper, we study the slightly stronger criterion of *signed edge recovery*: in particular, given a graphical model with parameter θ^*, we define the edge sign vector

$$E^* \quad := \quad \begin{cases} \text{sign}(\theta^*_{st}) & \text{if } (s,t) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note that the weaker graphical model selection problem amounts to recovering the vector $|E^*|$ of absolute values.

The classical notion of statistical consistency applies to the limiting behavior of an estimation procedure as the sample size n goes to infinity, with the model size p itself remaining fixed. In many contemporary applications of graphical models (e.g., gene microarrays, social networks etc.), the model dimension p is comparable or larger than the sample size n, so that the relevance of such "fixed p" asymptotics is doubtful. Accordingly, the goal of

this paper is to develop the broader notion of *high-dimensional consistency*, in which both the model dimension and the sample size are allowed to increase, and we study the scaling conditions under which consistent model selection is achievable.

More precisely, we consider sequences of graphical model selection problems, indexed by the sample size n, number of vertices p, and maximum node degree d. We assume that the sample size n goes to infinity, and both the problem dimension $p = p(n)$ and $d = d(n)$ may also scale as a function of n. The setting of fixed p or d is covered as a special case. Let $\widehat{E}_n$ be an estimator of the signed edge pattern E^*, based on the n samples. Our goal is to establish sufficient conditions on the scaling of the triple (n, p, d) such that our proposed estimator is consistent in the sense that

$$\mathbb{P}\left[\widehat{E}_n = E^*\right] \quad \rightarrow \quad 1 \qquad \text{as } n \rightarrow +\infty.$$

We sometimes call this property *sparsistency*, as a shorthand for consistency of the sparsity pattern of the parameter θ^*.

2.3 Neighborhood-based logistic regression

Note that recovering the signed edge vector E^* of an undirected graph G is equivalent to recovering, for each vertex $r \in V$, its *neighborhood set* $\mathcal{N}(r) := \{t \in V \mid (r,t) \in E\}$, along with the correct signs $\text{sign}(\theta^*_{rt})$ for all $t \in \mathcal{N}(r)$. To capture both the neighborhood structure and sign pattern, we define the *signed neighborhood set*

$$\mathcal{N}_{\pm}(r) \quad := \quad \{\text{sign}(\theta^*_{rt})\, t \mid t \in \mathcal{N}(s)\}\,. \tag{3}$$

The next step is to observe that this signed neighborhood set $\mathcal{N}_{\pm}(r)$ can be recovered from the sign-sparsity pattern of the $(p-1)$-dimensional subvector of parameters

$$\theta^*_{\backslash r} \quad := \quad \{\theta^*_{ru},\ u \in V \backslash r\}$$

associated with vertex r. In order to estimate this vector $\theta^*_{\backslash r}$, we consider the structure of the conditional distribution of X_r given the other variables $X_{\backslash r} = \{X_t \mid t \in V \backslash \{r\}\}$. A simple calculation shows that under the model (1), this conditional distribution takes the form

$$\mathbb{P}_{\theta^*}(x_r \mid x_{\backslash r}) \quad = \quad \frac{\exp\left(2x_r\left[\sum_{t \in V \backslash r} \theta^*_{rt} x_t\right]\right)}{\exp\left(2x_r[\sum_{t \in V \backslash r} \theta^*_{rt} x_t]\right) + 1}. \tag{4}$$

Thus, the variable X_r can be viewed as the response variable in a logistic regression in which all of the other variables $X_{\backslash r}$ play the role of the covariates.

With this set-up, our method for estimating the sign-sparsity pattern of the regression vector $\theta^*_{\backslash r}$ (and hence the neighborhood structure $\mathcal{N}_{\pm}(r)$) is based on computing the ℓ_1-regularized logistic regression of X_s on the other variables $X_{\backslash r}$. Explicitly, given a set of n i.i.d. samples $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, this regularized regression problem is a convex program, of the form $\min_{\theta_{\backslash r} \in \mathbb{R}^{p-1}} \left\{\ell(\theta; \{x^{(i)}\}) + \lambda_n \|\theta_{\backslash r}\|_1\right\}$, where $\lambda_n > 0$ is a regularization parameter,

to be specified by the user, and

$$\ell(\theta; \{x^{(i)}\}) \quad := \quad -\frac{1}{n} \sum_{i=1}^{n} \log \mathbb{P}_\theta(x_r^{(i)} \mid x_{\backslash r}^{(i)}) \tag{5}$$

is the rescaled negative log likelihood. (The rescaling factor $1/n$ in this definition is for later theoretical convenience.) Following some algebraic manipulation, the regularized negative log likelihood can be written as

$$\min_{\theta_{\backslash r} \in \mathbb{R}^{p-1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(\theta; x^{(i)}) - \sum_{u \in V \backslash r} \theta_{ru} \widehat{\mu}_{ru} + \lambda_n \|\theta_{\backslash r}\|_1 \right\}, \tag{6}$$

where

$$f(\theta; x) \quad := \quad \log \left(\exp(\sum_{t \in V \backslash r} \theta_{rt} x_t) + \exp(-[\sum_{t \in V \backslash r} \theta_{rt} x_t]) \right) \tag{7}$$

is a rescaled logistic loss, and $\widehat{\mu}_{ru} := \frac{1}{n} \sum_{i=1}^{n} x_r^{(i)} x_u^{(i)}$ are empirical moments. Note the objective function (6) is convex but not differentiable, due to the presence of the ℓ_1-regularizer. By Lagrangian duality, the problem (6) can be re-cast as a constrained problem over the ball $\|\theta_{\backslash r}\|_1 \leq C(\lambda_n)$. Consequently, by the Weierstrass theorem, the minimum over $\theta_{\backslash s}$ is always achieved.

Accordingly, let $\widehat{\theta}_{\backslash r}^n$ be an element of the minimizing set of problem (6). Although $\widehat{\theta}_{\backslash r}^n$ need not be unique in general since the problem (6) need not be strictly convex, our analysis shows that in the regime of interest, this minimizer $\widehat{\theta}_{\backslash r}^n$ is indeed unique. We use $\widehat{\theta}_{\backslash r}^n$ to estimate the signed neighborhood $\mathcal{N}_\pm(r)$ according to

$$\widehat{\mathcal{N}}_\pm(r) \quad := \quad \left\{ \text{sign}(\widehat{\theta}_{ru}) u \;\mid\; u \in V \backslash r, \; \widehat{\theta}_{su} \neq 0 \right\}. \tag{8}$$

We say that the full graph G is estimated consistently, written as the event $\{\widehat{G} = G(p, d)\}$, if $\widehat{\mathcal{N}}_\pm(r) = \mathcal{N}_\pm(r)$ for all $r \in V$.

3 Method and theoretical guarantees

Our main result concerns conditions on the sample size n relative to the parameters of the graphical model—more specifically, the number of nodes p and maximum node degree d—that ensure that the collection of signed neighborhood estimates (8), one for each node r of the graph, agree with the true neighborhoods, so that the full graph $G(p, d)$ is estimated consistently. In this section, we begin by stating the assumptions that underlie our main result, and then give a precise statement of the main result. We then provide a high-level overview of the key steps involved in its proof, deferring detail to later sections. Our analysis proceeds by first establishing sufficient conditions for correct signed neighborhood recovery—that is, $\{\widehat{\mathcal{N}}_\pm(r) = \mathcal{N}_\pm(r)\}$—for some fixed node $r \in V$. By showing that this neighborhood consistency is achieved at exponentially fast rates, we can then use a union bound over all p nodes of the graph to conclude that consistent graph selection $\{\widehat{G} = G(p, d)\}$ is also achieved.

3.1 Assumptions

Success of our method requires certain assumptions on the structure of the logistic regression problem. These assumptions are stated in terms of the Hessian of the likelihood function $\mathbb{E}\{\log \mathbb{P}_\theta[X_r \mid X_{\setminus r}]\}$, as evaluated at the true model parameter $\theta^*_{\setminus r} \in \mathbb{R}^{p-1}$. More specifically, for any fixed node $r \in V$, this Hessian is a $(p-1) \times (p-1)$ matrix of the form

$$Q^*_r \;:=\; \mathbb{E}_{\theta^*}\left\{\nabla^2 \log \mathbb{P}_{\theta^*}[X_r \mid X_{\setminus r}]\right\}. \tag{9}$$

For future reference, we calculate the explicit expression

$$Q^*_r \;=\; \mathbb{E}_{\theta^*}\left[\eta(X;\theta^*) X_{\setminus r} X^T_{\setminus r}\right] \tag{10}$$

where

$$\eta(u;\theta) \;:=\; \frac{4\exp\left(2u_r\left[\sum_{t\in V\setminus r}\theta_{rt}u_t\right]\right)}{(\exp\left(2u_r[\sum_{t\in V\setminus r}\theta_{rt}u_t]\right)+1)^2} \tag{11}$$

is the variance function. Note that the matrix Q^*_r is the Fisher information matrix associated with the local conditional probability distribution. Intuitively, it serves as the counterpart for discrete graphical models of the covariance matrix $\mathbb{E}[XX^T]$ of Gaussian graphical models, and indeed our assumptions are analogous to those imposed in previous work on the Lasso for Gaussian linear regression (Meinshausen and Bühlmann, 2006, Tropp, 2006, Zhao and Yu, 2007).

In the following we write simply Q^* for the matrix Q^*_r, where the reference node r should be understood implicitly. Moreover, we use $S := \{(r,t) \mid t \in \mathcal{N}(r)\}$ to denote the subset of indices associated with edges of r, and S^c to denote its complement. We use Q^*_{SS} to denote the $d \times d$ sub-matrix of Q^* indexed by S. With this notation, we state our assumptions:

[A1] Dependency condition: The subset of the Fisher information matrix corresponding to the relevant covariates has bounded eigenvalues: there exists a constant $C_{min} > 0$ such that

$$\Lambda_{min}(Q^*_{SS}) \;\geq\; C_{min}. \tag{12}$$

Moreover, we require that $\Lambda_{max}(\mathbb{E}_{\theta^*}[X_{\setminus r}X^T_{\setminus r}]) \leq D_{\max}$. These conditions ensure that the relevant covariates do not become overly dependent. (As stated earlier, we have suppressed notational dependence on r; thus this condition is assumed to hold for all $r \in V$.)

[A2] Incoherence condition: Our next assumption captures the intuition that the large number of irrelevant covariates (i.e., non-neighbors of node r) cannot exert an overly strong effect on the subset of relevant covariates (i.e., neighbors of node r). To formalize this intuition, we require the existence of an $\alpha \in (0,1]$ such that

$$|\!|\!| Q^*_{S^cS}(Q^*_{SS})^{-1} |\!|\!|_\infty \;\leq\; 1-\alpha. \tag{13}$$

3.2 Statement of main result

We are now ready to state our main result on the performance of neighborhood logistic regression for graphical model selection. Naturally, the limits of model selection are determined by the minimum value over the parameters θ^*_{rt} for pairs (r,t) included in the edge set of the true graph. Accordingly, we define the parameter

$$\theta^*_{\min} \quad = \quad \min_{(r,t)\in E} |\theta^*_{rt}|. \tag{14}$$

With this definition, we have the following

Theorem 1. *Consider a sequence of graphs $\{G(p,d)\}$ such that conditions A1 and A2 are satisfied by the population Fisher information matrices Q^*. If the sample size n satisfies*

$$n > Ld^3 \log(p) \tag{15}$$

for some constant L, and the minimum value $\theta^*_{\min}$ decays no faster than $\mathcal{O}(1/d)$, then for the regularization sequence $\lambda_n = 2\sqrt{\frac{\log p}{n}}$, the estimated graph $\widehat{G}(\lambda_n)$ obtained by neighborhood logistic regression satisfies

$$\mathbb{P}[\widehat{G}(\lambda_n) = G(p,d)] \quad = \quad \mathcal{O}\left(\exp\left(-K\frac{n}{d^3} - 3\log(p)\right)\right) \;\rightarrow\; 0 \tag{16}$$

for some constant K.

Remarks: For model selection in graphical models, one is typically interested in node degrees d that remain bounded (e.g., $d = \mathcal{O}(1)$), or that grow only weakly with graph size (say $d = o(p)$. In such cases, the growth condition (15) allows the number of observations to be substantially smaller than the graph size, i.e., the "large p, small n" regime. In particular, the graph size p can grow exponentially with the number of observations (i.e, $p(n) = \exp(n^\alpha)$ for some $\alpha \in (0,1)$.

In terms of the choice of regularization, the sequence λ_n needs to satisfy the following conditions:

$$n\lambda_n^2 > 2\log(p), \quad \text{and} \quad \sqrt{d}\lambda_n = \mathcal{O}(\theta^*_{\min}).$$

Under the growth condition (15), the choice $\lambda_n = 2\sqrt{\frac{\log p}{n}}$ suffices as long as $\theta^*_{\min}$ decays no faster than $\mathcal{O}(1/d)$.

The analysis required to prove Theorem 1 can be divided naturally into two parts. First, in Section 4, we prove a result (stated as Proposition 1) for "fixed design" matrices. More precisely, we show that if the dependence (A1) mutual incoherence (A2) conditions hold for the *sample Fisher information matrix*

$$Q^n \quad := \quad \widehat{\mathbb{E}}\left[\eta(X;\theta^*)X_{\backslash r}X_{\backslash r}^T\right] \;=\; \frac{1}{n}\sum_{i=1}^n \eta(x^{(i)};\theta^*)x_{\backslash r}^{(i)}(x_{\backslash r}^{(i)})^T \tag{17}$$

then the growth condition (15) and choice of λ_n from Theorem 1 are sufficient to ensure that the graph is recovered with high probability. Interestingly, our analysis shows that if the

conditions are imposed directly on the sample Fisher information matrices and $\theta^*_{\min} = \Theta(1)$, then the weaker growth condition $n = \Omega(d^2 \log(p))$ suffices for asymptotically exact graph recovery.

The second part of the analysis, provided in Section 5, is devoted to showing that under the specified growth conditions (A3), imposing incoherence and dependence assumptions on the *population version* of the Fisher information Q^* guarantees (with high probability) that analogous conditions hold for the sample quantities Q^n. While it follows immediately from the law of large numbers that the empirical Fisher information Q^n_{AA} converges to the population version Q^*_{AA} for any *fixed* subset A, the delicacy is that we require controlling this convergence over subsets of increasing size. The analysis therefore requires some large-deviations bounds, so as to provide exponential control on the rates of convergence.

3.3 Primal-dual witness for graph recovery

At a high-level, at the core of our proof lies the notion of a primal-dual witness. In particular, we explicitly construct an optimal *primal-dual pair*, namely, a primal solution $\widehat{\theta}$, along with an associated subgradient vector $\widehat{z}$ (which can be interpreted as a dual solution), such that the Karush-Kuhn-Tucker (KKT) conditions associated with the convex program (6) are satisfied. Moreover, we show that under the stated assumptions on (n, p, d), the primal-dual pair $(\widehat{\theta}, \widehat{z})$ can be constructed such that they act as a *witness*—that is, a certificate guaranteeing that the method correctly recovers the graph structure.

Let us write the convex program (6) in the form

$$\min_{\theta_{\setminus r} \in \mathbb{R}^{p-1}} \left\{ \ell(\theta; \{x^{(i)}\}) + \lambda_n \|\theta_{\setminus r}\|_1 \right\}, \tag{18}$$

where

$$\ell(\theta; \{x^{(i)}\}) \; = \; \ell(\theta) \; = \; \frac{1}{n} \sum_{i=1}^{n} f(\theta; x^{(i)}) - \sum_{u \in V \setminus r} \theta_{ru} \widehat{\mu}_{ru} \tag{19}$$

is the negative log likelihood associated with the logistic regression model. The KKT conditions associated with this model can be expressed as follows

$$\nabla \ell(\widehat{\theta}) + \lambda_n \widehat{z} \;\; = \;\; 0 \tag{20}$$

where the dual or subgradient vector $\widehat{z} \in \mathbb{R}^{p-1}$ satisfies the properties

$$\widehat{z}_{rt} = \text{sign}(\widehat{\theta}_{rt}) \quad \text{if } \widehat{\theta}_i \neq 0, \text{ and} \quad |\widehat{z}_{rt}| \leq 1 \quad \text{otherwise.} \tag{21}$$

One way to understand this vector $\widehat{z}$ is as a subgradient, meaning an element of the sub-differential of the ℓ_1-norm (see Rockafellar, 1970). An alternative interpretation is based on the constrained equivalent to problem (18), involving the constraint $\|\theta\|_1 \leq C(\lambda_n)$. This ℓ_1-constraint is equivalent to the family of constraints $\vec{v}^T \theta \leq C$, where the vector $\vec{v} \in \{-1, +1\}^{p-1}$ ranges over all possible sign vectors. In this formulation, the optimal dual vector is simply the conic combination

$$\widehat{z} = \sum_{\vec{v} \in \{-1,+1\}^{p-1}} \alpha^*_v \vec{v}, \tag{22}$$

where $\alpha_v^* \geq 0$ is the Lagrange multiplier associated with the constraint $\vec{v}^T\theta \leq C$.

The KKT conditions (20) and (21) must be satisfied by any optimal pair $(\widehat{\theta}, \widehat{z})$ to the convex program (18). In order for this primal-dual pair to correctly specify the graph structure, we require furthermore that the following properties are satisfied:

$$\begin{aligned} \operatorname{sign}(\widehat{z}_{rt}) &= \operatorname{sign}(\theta^*_{rt}) \qquad \text{for all } (r,t) \in S := \{(r,t) \mid t \in \mathcal{N}(r)\}, \qquad \text{and} && \text{(23a)} \\ \widehat{\theta}_{S^c} &= 0 \quad \text{where } S^c := \{(r,u) \mid (r,u) \notin E\}. && \text{(23b)} \end{aligned}$$

We now construct our *witness* pair $(\widehat{\theta}, \widehat{z})$ as follows. First, we set $\widehat{\theta}_S$ as the minimizer of the partial penalized likelihood,

$$\widehat{\theta}_S = \arg \min_{(\theta_S, 0) \in \mathbb{R}^{p-1}} \left\{ \ell(\theta; \{x^{(i)}\}) + \lambda_n \|\theta_S\|_1 \right\}, \tag{24}$$

and set $\widehat{z}_S = \operatorname{sign}(\widehat{\theta}_S)$. We then set $\widehat{\theta}_{S^c} = 0$ so that condition (23b) holds. Finally, we obtain $\widehat{z}_{S^c}$ from equation (20) by plugging in the values of $\widehat{\theta}$ and $\widehat{z}_S$. Thus, our construction satisfies conditions (23b) and (20). The remainder of the analysis consists of showing that our conditions on (n, p, d) imply that, with high-probability, the remaining conditions (23a) and (21) are satisfied.

This strategy is justified by the following lemma, which provides sufficient conditions for shared sparsity and uniqueness of the optimal solution:

Lemma 1 (Shared sparsity and uniqueness)**.** *If there exists an optimal primal solution $\widehat{\theta}$ with associated optimal dual vector $\widehat{z}$ such that $\|\widehat{z}_{S^c}\|_\infty < 1$, then any optimal primal solution $\widetilde{\theta}$ must have $\widetilde{\theta}_{S^c} = 0$. Moreover, if the Hessian sub-matrix $[\nabla^2 \ell(\widehat{\theta})]_{SS} \succ 0$, then $\widehat{\theta}$ is the unique optimal solution.*

Proof. By Lagrangian duality, the penalized problem (18) can be written as an equivalent constrained optimization problem over the ball $\|\theta\|_1 \leq C(\lambda_n)$, for some constant $C(\lambda_n) < +\infty$. Since the Lagrange multiplier associated with this constraint—namely, λ_n—is strictly positive, the constraint is active at any optimal solution, so that $\|\theta\|_1$ is constant across all optimal solutions. Consider the representation of $\widehat{z}$ as the convex combination (22) of sign vectors $\vec{v} \in \{-1, +1\}^{p-1}$, where the weights α_v^* are non-negative and sum to one. Since α^* is an optimal vector of Lagrange multipliers for the optimal primal solution $\widehat{\theta}$, it follows (Bertsekas, 1995) that any other optimal primal solution $\widetilde{\theta}$ must minimize the associated Lagrangian (i.e., satisfy equation (20)), and moreover must satisfy the complementary slackness conditions $\alpha_v^*[\vec{v}^T\theta - C] = 0$ for all sign vectors v. But these conditions imply that $\widehat{z}^T\widetilde{\theta} = C = \|\widetilde{\theta}\|_1$, which cannot occur if $\widetilde{\theta}_j \neq 0$ for some index j for which $|\widehat{z}_j| < 1$. We thus conclude that $\widetilde{\theta}_{S^c} = 0$ for all optimal primal solutions.

Finally, given that all optimal solutions satisfy $\theta_{S^c} = 0$, we may consider the restricted optimization problem subject to this set of constraints. If the principal submatrix of the Hessian is positive definite, then this sub-problem is strictly convex, so that the optimal solution must be unique. □

In our primal-dual witness proof, we exploit this lemma by constructing a primal-dual pair $(\widehat{\theta}, \widehat{z})$ such that $\|\widehat{z}_{S^c}\|_\infty < 1$. Moreover, under the conditions of Theorem 1, we prove that the sub-matrix of the sample Fisher information matrix is strictly positive definite with high probability, so that the primal solution $\widehat{\theta}$ is guaranteed to be unique.

4 Analysis under sample Fisher matrix assumptions

We begin by establishing model selection consistency when assumptions are imposed directly on the sample Fisher matrix Q^n, as opposed to on the population matrix Q^*, as in Theorem 1. In particular, we define the "good event"

$$\mathcal{M}(\{x^{(i)}\}) \;:=\; \Big\{\{x^{(i)}\} \mid Q^n = \widehat{\mathbb{E}}[\nabla^2 \ell(\theta^*)] \text{ satisfies } A1 \text{ and } A2\Big\}. \tag{25}$$

We then state the following

Proposition 1 (Consistency for fixed design)**.** *If $n > Ld^2 \log(p)$ for a suitably large constant L, and the minimum value $\theta^*_{\min}$ decays no faster than $\mathcal{O}(1/\sqrt{d})$, then for the regularization sequence $\lambda_n = 2\sqrt{\frac{\log p}{n}}$, the estimated graph $\widehat{G}(\lambda_n)$ obtained by neighborhood logistic regression satisfies*

$$\mathbb{P}[\widehat{G}(\lambda_n) = G(p,d) \mid \mathcal{M}(\{x^{(i)}\})] \;=\; \mathcal{O}\left(\exp\left(-n\lambda_n^2 - \log(p)\right)\right) \;\to\; 0. \tag{26}$$

Loosely stated, this result guarantees that if the sample Fisher information matrix is "good", then the conditional probability of successful graph recovery converges to zero at the specified rate. The remainder of this section is devoted to the proof of Proposition 1.

4.1 Key technical results

We begin with statements of some key technical lemmas that are central to our main argument, with their proofs deferred to Appendix A. The central object is the following expansion, obtained by re-writing the zero-subgradient condition as

$$\nabla \ell(\widehat{\theta}; \{x^{(i)}\}) - \nabla \ell(\theta^*; \{x^{(i)}\}) \;=\; W^n - \lambda_n \widehat{z}, \tag{27}$$

where we have introduced the short-hand notation W^n for the $(p-1)$-vector

$$\begin{aligned} W^n := -\nabla \ell(\theta^*; \{x^{(i)}\}) \;&=\; \\ -\frac{1}{n}\sum_{i=1}^n x^{(i)}_{\setminus r} &\left\{ x^{(i)}_r - \frac{\exp(\sum_{t \in V \setminus r} \theta^*_{rt} x^{(i)}_t) - \exp(-[\sum_{t \in V \setminus r} \theta^*_{rt} x^{(i)}_t])}{\exp(\sum_{t \in V \setminus r} \theta^*_{rt} x^{(i)}_t) + \exp(-[\sum_{t \in V \setminus r} \theta^*_{rt} x^{(i)}_t])} \right\} \\ =\; -\frac{1}{n}\sum_{i=1}^n x^{(i)}_{\setminus r} &\left\{ x^{(i)}_r - \mathbb{P}_{\theta^*}[x_r = 1 \mid x^{(i)}_{\setminus r}] + \mathbb{P}_{\theta^*}[x_r = -1 \mid x^{(i)}_{\setminus r}] \right\}. \end{aligned}$$

For future reference, note that $\mathbb{E}_{\theta^*}[W^n] = 0$. Next, applying the mean-value theorem coordinate-wise to the expansion (27) yields

$$\nabla^2 \ell(\theta^*; \{x^{(i)}\})\, [\widehat{\theta} - \theta^*] \;=\; W^n - \lambda_n \widehat{z} + R^n, \tag{28}$$

where the remainder term takes the form

$$R^n_j \;=\; \left[\nabla^2 \ell(\bar{\theta}^{(j)}; \{x^{(i)}\}) - \nabla^2 \ell(\theta^*; \{x^{(i)}\})\right]^T_j (\widehat{\theta} - \theta^*), \tag{29}$$

with $\bar{\theta}^{(j)}$ a parameter vector on the line between θ^* and $\widehat{\theta}$, and with $[\cdot]^T_j$ denoting the jth row of the matrix. The following lemma addresses the behavior of the term W^n in this expansion:

Lemma 2. *If $n\lambda_n^2 > \log(p)$, then for the specified mutual incoherence parameter $\alpha \in (0,1]$, we have*

$$\mathbb{P}\left(\frac{2-\alpha}{\lambda_n}\|W^n\|_\infty \geq \frac{\alpha}{4}\right) \quad = \quad \mathcal{O}\left(\exp\left(-n\lambda_n^2 + \log(p)\right)\right) \to 0. \tag{30}$$

See Appendix A.1 for the proof of this claim.

The following lemma establishes that the sub-vector $\widehat{\theta}_S$ is an ℓ_2-consistent estimate of the true sub-vector θ^*_S:

Lemma 3 (ℓ_2-consistency of primal subvector)**.** *If $\lambda_n d \leq \frac{C_{min}^2}{10 D_{\max}}$, then as $n \to +\infty$, we have*

$$\|\widehat{\theta}_S - \theta^*_S\|_2 = \mathcal{O}_p\left(\sqrt{d}\lambda_n\right) \to 0. \tag{31}$$

See Appendix A.2 for the proof of this claim.

Our final technical lemma provides control on the the remainder term (29):

Lemma 4. *If $n\lambda_n^2 > \log(p)$ and $d\lambda_n$ is sufficiently small, then for mutual incoherence parameter $\alpha \in (0,1]$, we have*

$$\mathbb{P}\left(\frac{2-\alpha}{\lambda_n}\|R^n\|_\infty \geq \frac{\alpha}{4}\right) \quad = \quad \mathcal{O}(\exp\left(-n\lambda_n^2 + \log(p)\right)) \to 0. \tag{32}$$

See Appendix A.3 for the proof of this claim.

4.2 Proof of Proposition 1

Using these lemmas, we can now complete the proof of Proposition 1. Recalling our shorthand $Q^n = \nabla^2_\theta \ell(\theta^*; \{x^{(i)}\})$, we re-write condition (28) in block form as:

$$Q^n_{S^cS}\left[\widehat{\theta}_S - \theta^*_S\right] \quad = \quad W^n_{S^c} - \lambda_n \widehat{z}_{S^c} + R^n_{S^c}, \tag{33a}$$

$$Q^n_{SS}\left[\widehat{\theta}_S - \theta^*_S\right] \quad = \quad W^n_S - \lambda_n \widehat{z}_S + R^n_S. \tag{33b}$$

Since the matrix Q^n_{SS} is invertible by assumption, the conditions (33) can be re-written as

$$Q^n_{S^cS}\left(Q^n_{SS}\right)^{-1}\left[W^n_S - \lambda_n \widehat{z}_S + R^n_S\right] \quad = \quad W^n_{S^c} - \lambda_n \widehat{z}_{S^c} + R^n_{S^c}. \tag{34}$$

Rearranging yields the condition

$$\left[W^n_{S^c} - R^n_{S^c}\right] - Q^n_{S^cS}\left(Q^n_{SS}\right)^{-1}\left[W^n_S - R^n_S\right] + \lambda_n Q^n_{S^cS}\left(Q^n_{SS}\right)^{-1}\widehat{z}_S \quad = \quad \lambda_n \widehat{z}_{S^c}. \tag{35}$$

Strict dual feasibility We now demonstrate that for the dual sub-vector $\widehat{z}_{S^c}$ defined by equation (35), we have $\|\widehat{z}_{S^c}\|_\infty < 1$. Using the triangle inequality and the mutual incoherence bound (13), we have that

$$\begin{aligned}
\|\widehat{z}_{S^c}\|_\infty &\leq \left|\!\left|\!\left| Q^n_{S^cS}\left(Q^n_{SS}\right)^{-1}\right|\!\right|\!\right|_\infty \left[\frac{\|W^n_S\|_\infty}{\lambda_n} + \frac{\|R^n_S\|_\infty}{\lambda_n} + 1\right] + \frac{\|R^n_{S^c}\|_\infty}{\lambda_n} + \frac{\|W^n_{S^c}\|_\infty}{\lambda_n} \\
&\leq (1-\alpha) + (2-\alpha)\left[\frac{\|R^n\|_\infty}{\lambda_n} + \frac{\|W^n\|_\infty}{\lambda_n}\right]
\end{aligned} \tag{36}$$

Next, applying Lemmas 2 and 4, we have

$$\|\widehat{z}_{S^c}\|_\infty \;\leq\; (1-\alpha) + \frac{\alpha}{4} + \frac{\alpha}{4} \;=\; 1 - \frac{\alpha}{2}.$$

with probability converging to one.

Correct sign recovery: We next show that our primal sub-vector $\widehat{\theta}_S$ defined by equation (24) satisfies sign consistency, meaning, $\mathrm{sgn}(\widehat{\theta}_S) = \mathrm{sgn}(\theta^*_S)$. In order to establish this, it suffices to show that

$$\|\theta_S - \theta^*_S\|_\infty \;\leq\; \frac{\theta^*_{\min}}{2}$$

where we recall the notation $\theta^*_{\min} := \min_{(r,t)\in E} |\theta^*_{rt}|$. From Lemma 3, we have $\|\theta_S - \theta^*_S\|_2 = \mathcal{O}_p(\sqrt{d}\lambda_n)$, so that

$$\frac{2}{\theta^*_{\min}}\|\theta_S - \theta^*_S\|_\infty \;\leq\; \frac{2}{\theta^*_{\min}}\|\theta_S - \theta^*_S\|_2 \tag{37}$$

$$= \;\mathcal{O}\left(\frac{\sqrt{d}\lambda_n}{\theta^*_{\min}}\right) \tag{38}$$

Since $\theta^*_{\min}$ decays no faster than $\Theta(1/\sqrt{d})$, the right-hand side is upper bounded by $\mathcal{O}(\lambda_n d)$, which can be made smaller than 1 by choosing λ_n sufficiently small, as asserted in Proposition 1.

5 Uniform convergence of sample information matrices

In this section, we complete the proof of Theorem 1 by showing that if the dependency ($A1$) and incoherence ($A2$) assumptions are imposed on the *population* Fisher information matrix then under the specified scaling of (n, p, d), analogous bounds hold for the *sample* Fisher information matrices with probability converging to one. These results are not immediate consequences of classical random matrix theory (e.g., Davidson and Szarek (2001)), since the elements of Q^n are highly dependent.

Recall the definitions

$$Q^* := \mathbb{E}_{\theta^*}\left[\eta(X;\theta^*)X_{\setminus r}X^T_{\setminus r}\right], \quad \text{and} \quad Q^n := \widehat{\mathbb{E}}\left[\eta(X;\theta^*)X_{\setminus r}X^T_{\setminus r}\right], \tag{39}$$

where $\mathbb{E}_{\theta^*}$ denotes the population expectation, and $\widehat{\mathbb{E}}$ denotes the empirical expectation, and the variance function η was defined previously equation (11). The following lemma asserts the eigenvalue bounds in Assumption $A1$ hold with high probability for sample covariance matrices:

Lemma 5. *Suppose that assumption $A1$ holds for the population matrix Q^* and $\mathbb{E}_{\theta^*}[XX^T]$. For any $\delta > 0$ and some fixed constants A and B, we have*

$$\mathbb{P}\left[\Lambda_{max}\big[\frac{1}{n}\sum_{i=1}^n x^{(i)}_{\setminus r}(x^{(i)}_{\setminus r})^T\big] \geq D_{\max} + \delta\right] \;\leq\; 2\exp\left(-A\frac{\delta^2 n}{d^2} + B\log(d)\right). \tag{40a}$$

$$\mathbb{P}[\Lambda_{min}(Q^n_{SS}) \leq C_{min} - \delta] \;\leq\; 2\exp\left(-A\frac{\delta^2 n}{d^2} + B\log(d)\right). \tag{40b}$$

The following result is the analog for the incoherence assumption ($A2$), showing that the scaling of (n, p, d) given in Theorem 1 guarantees that population incoherence implies sample incoherence.

Lemma 6. *If the population covariance satisfies a mutual incoherence condition* (13) *with parameter* $\alpha \in (0, 1]$ *as in Assumption A2, then the sample matrix satisfies an analogous version, with high probability, in the sense that*

$$\mathbb{P}\left[\|\!|Q^n_{S^cS}(Q^n_{SS})^{-1}|\!\|_\infty \geq 1 - \frac{\alpha}{2}\right] \;\leq\; \exp\left(-K\frac{n}{d^3} + \log(p)\right). \tag{41}$$

Proofs of these two lemmas are provided in the following sections. Before proceeding, we begin by taking note of a simple bound to be used repeatedly throughout our arguments. By definition of the matrices $Q^n(\theta)$ and $Q(\theta)$ (see equations (17) and (10)), the $(j,k)^{th}$ element of the difference matrix $Q^n(\theta) - Q(\theta)$ can be written as an i.i.d. sum of the form $Z_{jk} = \frac{1}{n}\sum_{i=1}^n Z^{(i)}_{jk}$, where each $Z^{(i)}_{jk}$ is zero-mean and bounded (in particular, $|Z^{(i)}_{jk}| \leq 4$). By the Azuma-Hoeffding bound (Hoeffding, 1963), for any indices $j, k = 1, \ldots, d$ and for any $\epsilon > 0$, we have

$$\mathbb{P}[(Z_{jk})^2 \geq \epsilon^2] \;=\; \mathbb{P}[|\frac{1}{n}\sum_{i=1}^n Z^{(i)}_{jk}| \geq \epsilon] \;\leq\; 2\exp\left(-\frac{\epsilon^2 n}{32}\right). \tag{42}$$

So as to simplify notation, throughout this section, we use K to denote a universal positive constant, independent of (n, p, d). Note that the precise value and meaning of K may differ from line to line.

5.1 Proof of Lemma 5

By the Courant-Fischer variational representation (Horn and Johnson, 1985), we have

$$\begin{aligned}
\Lambda_{min}(Q_{SS}) &= \min_{\|x\|_2=1} x^T Q_{SS} x \\
&= \min_{\|x\|_2=1} \left\{x^T Q^n_{SS} x + x^T(Q_{SS} - Q^n_{SS})x\right\} \\
&\leq y^T Q^n_{SS} y + y^T(Q_{SS} - Q^n_{SS})y,
\end{aligned}$$

where $y \in \mathbb{R}^d$ is a unit-norm minimal eigenvector of Q^n_{SS}. Therefore, we have

$$\Lambda_{min}(Q^n_{SS}) \;\geq\; \Lambda_{min}(Q_{SS}) - |\!\|Q_{SS} - Q^n_{SS}|\!\|_2 \;\geq\; C_{min} - |\!\|Q_{SS} - Q^n_{SS}|\!\|_2.$$

Hence it suffices to obtain a bound on the spectral norm $|\!\|Q_{SS} - Q^n_{SS}|\!\|_2$. Observe that

$$|\!\|Q^n_{SS} - Q_{SS}|\!\|_2 \;\leq\; \big[\sum_{j=1}^d \sum_{k=1}^d (Z_{jk})^2\big]^{\frac{1}{2}}.$$

Setting $\epsilon^2 = \delta^2/d^2$ in equation (42) and applying the union bound over the d^2 index pairs (j, k) then yields

$$\mathbb{P}[|\!\|Q^n_{SS} - Q_{SS}|\!\|_2 \geq \delta] \;\leq\; 2\exp\left(-K\frac{\delta^2 n}{d^2} + 2\log(d)\right). \tag{43}$$

Similarly, we have

$$\mathbb{P}[\Lambda_{max}(\frac{1}{n}\sum_{i=1}^{n} x_{\backslash r}^{(i)}(x_{\backslash r}^{(i)})^T) \geq D_{\max}] \quad \leq \quad \mathbb{P}\left[\|\!| (\frac{1}{n}\sum_{i=1}^{n} x_{\backslash r}^{(i)}(x_{\backslash r}^{(i)})^T) - \mathbb{E}_{\theta^*}[X_{\backslash r} X_{\backslash r}^T] |\!\|_2 \geq \delta \right],$$

which obeys the same upper bound (43), by following the analogous argument.

5.2 Proof of Lemma 6

We begin by decomposing the sample matrix as the sum $Q^n_{S^cS}(Q^n_{SS})^{-1} = T_1 + T_2 + T_3 + T_4$, where we define

$$\begin{aligned}
T_1 &:= Q^*_{S^cS}\left[(Q^n_{SS})^{-1} - (Q^*_{SS})^{-1}\right] && \text{(44a)}\\
T_2 &:= [Q^n_{S^cS} - Q^*_{S^cS}]\,(Q^*_{SS})^{-1} && \text{(44b)}\\
T_3 &:= [Q^n_{S^cS} - Q^*_{S^cS}]\left[(Q^n_{SS})^{-1} - (Q^*_{SS})^{-1}\right] && \text{(44c)}\\
T_4 &:= Q^*_{S^cS}(Q^*_{SS})^{-1} && \text{(44d)}
\end{aligned}$$

The fourth term is easily controlled; indeed, we have $|\!|\!| T_4 |\!|\!|_\infty = |\!|\!| Q^*_{S^cS}(Q^*_{SS})^{-1} |\!|\!|_\infty \leq 1-\alpha$ by the incoherence assumption $A2$. If we can show that $|\!|\!| T_i |\!|\!|_\infty \leq \frac{\alpha}{6}$ for the remaining indices $i = 1, 2, 3$, then by our four term decomposition and the triangle inequality, the sample version satisfies the bound (41), as claimed. We deal with these remaining terms using the following lemmas:

Lemma 7. *For any $\delta > 0$ and constants K, K', the following bounds hold:*

$$\begin{aligned}
\mathbb{P}[|\!|\!| Q^n_{S^cS} - Q^*_{S^cS} |\!|\!|_\infty \geq \delta] &\leq 2\exp\left(-K\frac{n\,\delta^2}{d^2} + \log(d) + \log(p-d)\right) && \text{(45a)}\\
\mathbb{P}[|\!|\!| Q^n_{SS} - Q^*_{SS} |\!|\!|_\infty \geq \delta] &\leq 2\exp\left(-K\frac{n\,\delta^2}{d^2} + 2\log(d)\right) && \text{(45b)}\\
\mathbb{P}[|\!|\!| (Q^n_{SS})^{-1} - (Q^*_{SS})^{-1} |\!|\!|_\infty \geq \delta] &\leq 4\exp\left(-K\frac{n\,\delta^2}{d^3} + K'\log(d)\right). && \text{(45c)}
\end{aligned}$$

See Appendix B for the proof of these claims.

Control of first term: Turning to the first term, we first re-factorize it as

$$T_1 \quad = \quad Q^*_{S^cS}(Q^*_{SS})^{-1}\,[Q^n_{SS} - Q^*_{SS}]\,(Q^n_{SS})^{-1},$$

and then bound it (using the sub-multiplicative property $|\!|\!| AB |\!|\!|_\infty \leq |\!|\!| A |\!|\!|_\infty |\!|\!| B |\!|\!|_\infty$) as follows

$$\begin{aligned}
|\!|\!| T_1 |\!|\!|_\infty &\leq |\!|\!| Q^*_{S^cS}(Q^*_{SS})^{-1} |\!|\!|_\infty |\!|\!| Q^n_{SS} - Q^*_{SS} |\!|\!|_\infty |\!|\!| (Q^n_{SS})^{-1} |\!|\!|_\infty\\
&\leq (1-\alpha)\,|\!|\!| Q^n_{SS} - Q^*_{SS} |\!|\!|_\infty \left\{\sqrt{d}\,|\!|\!| (Q^n_{SS})^{-1} |\!|\!|_2\right\},
\end{aligned}$$

where we have used the incoherence assumption $A2$. Using the bound (40b) from Lemma 5 with $\delta = C_{min}/2$, we have $|\!|\!| (Q^n_{SS})^{-1} |\!|\!|_2 = [\Lambda_{min}(Q^n_{SS})]^{-1} \leq \frac{2}{C_{min}}$ with probability greater

than $1 - \exp\left(-Kn/d^2 + 2\log(d)\right)$. Next, applying the bound (45b) with $\delta = c/\sqrt{d}$, we conclude that with probability greater than $1 - 2\exp\left(-Knc^2/d^3 + \log(d)\right)$, we have

$$\|\!|Q^n_{SS} - Q^*_{SS}|\!\|_\infty \;\leq\; c/\sqrt{d}.$$

By choosing the constant $c > 0$ sufficiently small, we are guaranteed that

$$\mathbb{P}[\|\!|T_1|\!\|_\infty \geq \alpha/6] \;\leq\; 2\exp\left(-K\frac{nc^2}{d^3} + \log(d)\right). \tag{46}$$

Control of second term: To bound T_2, we first write

$$\begin{aligned} \|\!|T_2|\!\|_\infty &\leq \sqrt{d}\|\!|(Q^*_{SS})^{-1}|\!\|_2 \, \|\!|Q^n_{S^cS} - Q^*_{S^cS}|\!\|_\infty \\ &\leq \frac{\sqrt{d}}{C_{min}} \, \|\!|Q^n_{S^cS} - Q^*_{S^cS}|\!\|_\infty \end{aligned}$$

We then apply bound (45a) with $\delta = \frac{\alpha}{3}\frac{C_{min}}{\sqrt{d}}$ to conclude that

$$\mathbb{P}[\|\!|T_2|\!\|_\infty \geq \alpha/3] \;\leq\; 2\exp\left(-K\frac{n}{d^3} + \log(p-d)\right). \tag{47}$$

Control of third term: Finally, in order to bound the third term T_3, we apply the bounds (45a) and (45b), both with $\delta = \sqrt{\alpha/3}$, and use the fact that $\log(d) \leq \log(p-d)$ to conclude that

$$\mathbb{P}[\|\!|T_3|\!\|_\infty \geq \alpha/3] \;\leq\; 4\exp\left(-K\frac{n}{d^3} + \log(p-d)\right). \tag{48}$$

Putting together all of the pieces, we conclude that

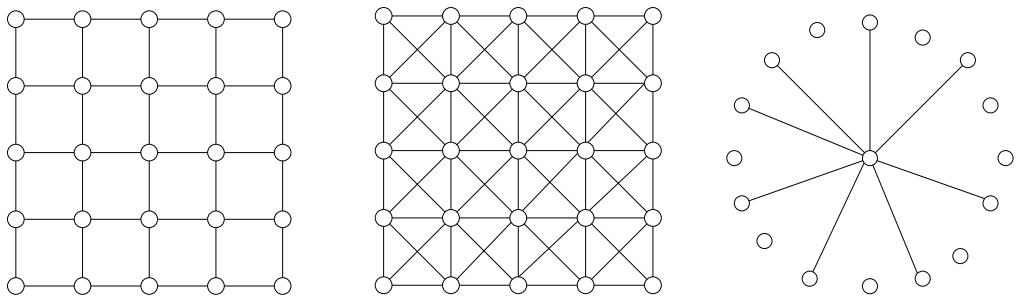
$$\mathbb{P}[\|\!|Q^n_{S^cS}(Q^n_{SS})^{-1}|\!\|_\infty \geq 1 - \alpha/2] \;=\; \mathcal{O}\left(\exp(-K\frac{n}{d^3} + \log(p))\right).$$

as claimed.

6 Experimental results

We now describe experimental results that illustrate some consequences of Theorem 1, for various types of graphs and scalings of (n, p, d). In all cases, we solved the ℓ_1-regularized logistic regression using special purpose interior-point code developed and described by Koh et al. (2007).

We performed experiments for three different classes of graphs: four-nearest neighbor lattices, (b) eight-nearest neighbor lattices, and (c) star-shaped graphs, as illustrated in Figure 1. Given a distribution $\mathbb{P}_{\theta^*}$ of the Ising form (1), we generated random data sets $\{x^{(1)}, \ldots, x^{(n)}\}$ by Gibbs sampling for the lattice models, and by exact sampling for the star graph. For a given graph class and edge strength $\omega > 0$, we examined the performance of models with *mixed couplings*, meaning $\theta^*_{st} = \pm\omega$ with equal probability, or with *positive*



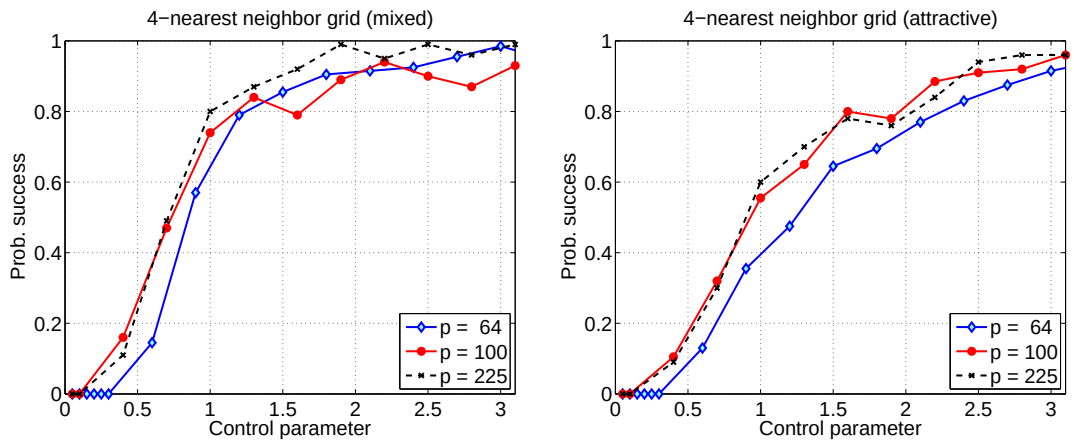
(a) (b) (c)

Figure 1. Illustrations of different graph classes used in simulations. (a) Four-nearest neighbor grid ($d = 4$). (b) Eight-nearest neighbor grid ($d = 8$). (c) Star-shaped graph ($d = \Theta(p)$, or $d = \Theta(\log(p))$).

couplings, meaning that $\theta^*_{st} = \omega$ for all edges (s, t). In all cases, we set the regularization parameter as $\lambda_n = \Theta(\sqrt{\frac{\log p}{n}})$. Above the threshold sample size n predicted by Theorem 1, this choice ensured correct model selection with high probability, consistent with the theoretical prediction. For any given graph and coupling type, we performed simulations for sample sizes n scaling as $n = 10\beta d \log(p)$, where the control parameter β ranged from 0.1 to upwards of 2, depending on the graph type.

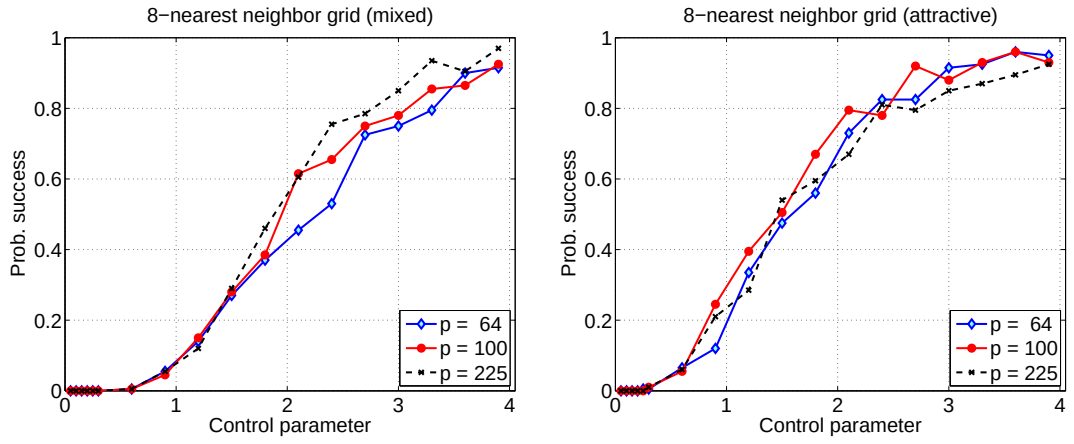
Figure 2 shows results for the 4-nearest-neighbor grid model, illustrated in Figure 1(a), for three different graph sizes $p \in \{64, 100, 225\}$, with mixed couplings (panel (a)) and attractive couplings (panel (b)). Each curve corresponds to a given problem size, and corresponds to the success probability versus the control parameter β. Each point corresponds to the average of $N = 200$ trials. Notice how despite the very different regimes of (n, p) that underlie each curve, the different curves all line up with one another quite well. This fact shows that for a fixed degree graph (in this case deg $= 4$), the ratio $n/\log(p)$ controls the success/failure of our model selection procedure, consistent with the prediction of Theorem 1. Figure 3 shows analogous results for the 8-nearest-neighbor lattice model ($d = 8$), for the same range of problem size $p \in \{64, 100, 225\}$, as well as both mixed and attractive couplings. Notice how once again the curves for different problem sizes are all well-aligned, consistent with the prediction of Theorem 1.

For our last set of experiments, we investigated the performance of our method for a class of graphs with unbounded maximum degree d. In particular, we constructed star-shaped graphs with p vertices by designating one node as the spoke, and connecting it to $d < (p-1)$ of its neighbors. For linear sparsity, we chose $d = \lceil 0.1p \rceil$, whereas for logarithmic sparsity we choose $d = \lceil \log(p) \rceil$. We again studied a triple of graph sizes $p \in \{64, 100, 225\}$ and Figure 4 shows the resulting curves of success probability versus control parameter $\beta = n/[10d \log(p)]$. Panels (a) and (b) correspond respectively to the cases of logarithmic and linear degrees. As with the bounded degree models in Figure 2 and 3, these curves align with one another, showing a transition from failure to success with probability one.



(a) (b)

Figure 2. Plots of success probability $\mathbb{P}[\widehat{\mathcal{N}}_{\pm}(r) = \mathcal{N}(r), \forall r]$ versus the control parameter $\beta(n, p, d) = n/[10d\log(p)]$ for Ising models on 2-D grids with four nearest-neighbor interactions ($d = 4$). (a) Randomly chosen mixed sign couplings $\theta^*_{st} = \pm 0.50$. (b) All positive couplings $\theta^*_{st} = 0.50$.



(a) (b)

Figure 3. Plots of success probability $\mathbb{P}[\widehat{\mathcal{N}}_{\pm}(r) = \mathcal{N}_{\pm}(r)]$ versus the control parameter $\beta(n, p, d) = n/[10d\log(p)]$ for Ising models on 2-D grids with eight nearest-neighbor interactions ($d = 8$). (a) Randomly chosen mixed sign couplings $\theta^*_{st} = \pm 0.25$. (b) All positive couplings $\theta^*_{st} = 0.25$.

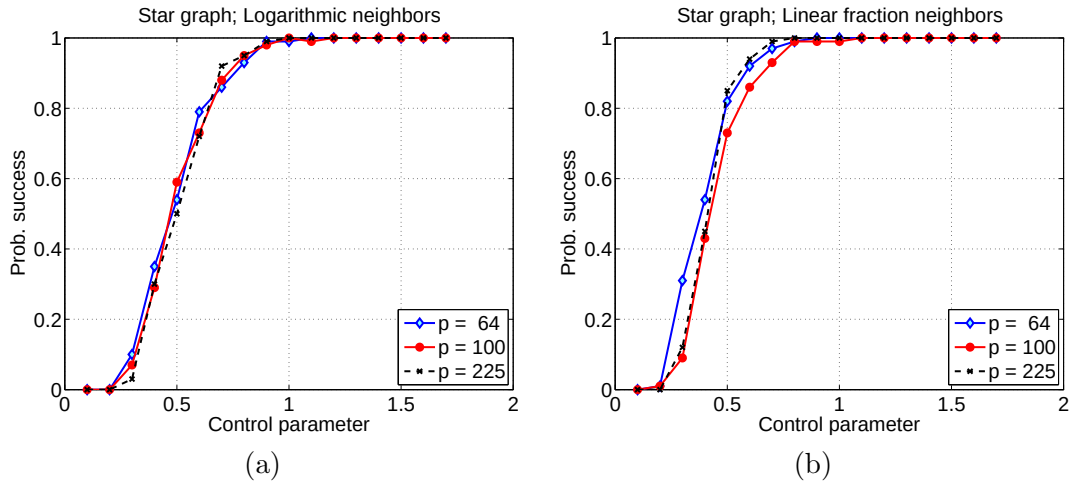


Figure 4. Plots of success probability $\widehat{\mathcal{N}}_{\pm}(r) = \mathcal{N}_{\pm}(r)]$ versus the control parameter $\beta(n, p, d) = n/[10d\log(p)]$ for star-shaped graphs, in which $d = \Theta(p)$, for attractive couplings. (a) Logarithmic growth in degrees. (b) Linear growth in degrees.

7 Conclusion

We have shown that a technique based on ℓ_1-regularized logistic regression can be used to perform consistent model selection in discrete graphical models, with polynomial computational complexity and sample complexity logarithmic in the graph size. Our analysis applies to the high-dimensional setting, in which both the number of nodes p and maximum neighborhood sizes d are allowed to grow as a function of the number of observations n. There are a number of possible directions for future work. For bounded degree graphs, our results show that the structure can be recovered with high probability once $n/\log(p)$ is sufficiently large. Up to constant factors, this result matches known information-theoretic lower bounds (Bresler et al., 2008, Santhanam and Wainwright, 2008). On the other hand, our experimental results on graphs with growing degrees (star-shaped graphs) are consistent with the conjecture that the logistic regression procedure exhibits a threshold at a sample size $n = \Theta(d \log p)$, at least for problems where the minimum value $\theta^*_{\min}$ stays bounded away from zero. It would be interesting to provide a sharp threshold result for this problem, to parallel the known thresholds for ℓ_1-regularized linear regression, or the Lasso (see Wainwright (2006)). Finally, the ideas described here, while specialized in this paper to the pairwise binary case, are more broadly applicable to discrete graphical models with a higher number of states; this is an interesting direction for future research.

Acknowledgements

Research supported in part by NSF grants IIS-0427206 and CCF-0625879 (PR and JL), NSF grants DMS-0605165 and CCF-0545862 (PR and MJW), and a Siebel Scholarship (PR).

A Proofs for Section 4.1

In this section, we provide proofs of Lemmas 2, Lemma 3 and Lemma 4, previously stated in Section 4.1.

A.1 Proof of Lemma 2

Note that any entry of W^n has the form $W_u^n = \frac{1}{n}\sum_{i=1}^n Z_u^{(i)}$, where for $i = 1, 2, \ldots, n$, the variables

$$Z_u^{(i)} \quad := \quad x_{\setminus r}^{(i)} \left\{ x_r^{(i)} - \mathbb{P}_{\theta^*}[x_r = 1 \mid x_{\setminus r}^{(i)}] + \mathbb{P}_{\theta^*}[x_r = -1 \mid x_{\setminus r}^{(i)}] \right\}$$

are zero-mean under $\mathbb{P}_{\theta^*}$, i.i.d., and bounded ($|Z_u^{(i)}| \leq 2$). Therefore, by the Azuma-Hoeffding inequality (Hoeffding, 1963), we have, for any $\delta > 0$, $\mathbb{P}\left[|W_u^n| > \delta\right] \leq 2\exp\left(-\frac{n\delta^2}{8}\right)$. Setting $\delta = \frac{\alpha\lambda_n}{4\,(2-\alpha)}$, we obtain

$$\mathbb{P}\left[\frac{2-\alpha}{\lambda_n}|W_u^n| > \frac{\alpha}{4}\right] \quad \leq \quad 2\exp\left(-Kn\lambda_n^2\right)$$

for some constant K. Finally, applying a union bound over the indices u of W^n yields

$$\mathbb{P}\left[\frac{2-\alpha}{\lambda_n}\|W^n\|_\infty > \frac{\alpha}{4}\right] \quad \leq \quad 2\exp\left(-Kn\lambda_n^2 + \log(p)\right),$$

as claimed.

A.2 Proof of Lemma 3

Following a method of Rothman et al. (2008), we define the function $G : \mathbb{R}^d \to \mathbb{R}$ by

$$G(u_S) \quad := \quad \ell(\theta_S^* + u_S; \{x^{(i)}\}) - \ell(\theta_S^*; \{x^{(i)}\}) + \lambda_n\left(\|\theta_S^* + u_S\| - \|\theta_S^*\|\right). \tag{49}$$

It can be seen from equation (24) that $\widehat{u} = \widehat{\theta}_S - \theta_S^*$ minimizes G. Note also that G is convex, and moreover $G(0) = 0$ by construction; therefore, we must have $G(\widehat{u}) \leq 0$. Suppose that we show that for some radius $B > 0$, and for $u \in \mathbb{R}^d$ with $\|u\|_2 = B$, we have $G(u) > 0$. We then claim that $\|\widehat{u}\|_2 \leq B$. Indeed, if $\widehat{u}$ lay outside the ball of radius B, then the convex combination $t\widehat{u} + (1-t)(0)$ would lie on the boundary of the ball, for an appropriately chosen $t \in (0, 1)$. By convexity,

$$G\left(t\widehat{u} + (1-t)(0)\right) \;\leq\; tG(\widehat{u}) + (1-t)G(0) \;\leq\; 0,$$

contradicting the assumed strict positivity of G on the boundary.

It thus suffices to establish strict positivity of G on the boundary of the ball with radius $B = M\lambda_n\sqrt{d}$, where $M > 0$ is a parameter to be chosen later in the proof. Let $u \in \mathbb{R}^d$ be an arbitrary vector with $\|u\|_2 = B$. Recalling the notation $W = \nabla\ell(\theta^*; \{x^{(i)}\})$, by a Taylor series expansion of the log likelihood component of G, we have

$$G(u) \quad = \quad W_S^T u + u^T\left[\nabla^2\ell(\theta_S^* + \alpha u)\right]u + \lambda_n\left(\|\theta_S^* + u_S\| - \|\theta_S^*\|\right), \tag{50}$$

for some $\alpha \in [0,1]$. For the first term, we have the bound

$$|W_S^T u| \;\leq\; \|W_S\|_\infty \|u\|_1 \;\leq\; \|W_S\|_\infty \sqrt{d}\,\|u\|_2 \;\leq\; \left(\lambda_n \sqrt{d}\right)^2 \frac{M}{4}, \tag{51}$$

since $\|W_S\|_\infty \leq \frac{\lambda_n}{4}$ with probability converging to one from Lemma 2.

Applying the triangle inequality to the last term in the expansion (50) yields

$$\lambda_n \|\theta_S^* + u_S\|_1 - \|\theta_S^*\|_1 \;\geq\; -\lambda_n \|u_S\|_1 \;\geq\; -\lambda_n \sqrt{d} \|u_S\|_2 \;=\; -M\left(\sqrt{d}\lambda_n\right)^2. \tag{52}$$

Finally, turning to the middle Hessian term, we have

$$\begin{aligned} q^* := \Lambda_{min}(\nabla^2 \ell(\theta_S^* + \alpha u; \{x^{(i)}\})) \;&\geq\; \min_{\alpha \in [0,1]} \Lambda_{min}(\nabla^2 \ell(\theta_S^* + \alpha u_S; \{x^{(i)}\})) \\ &=\; \min_{\alpha \in [0,1]} \Lambda_{min}\left[\frac{1}{n}\sum_{i=1}^n \eta(x^{(i)}; \theta_S^* + \alpha u_S) x_S^{(i)} (x_S^{(i)})^T\right] \end{aligned}$$

By a Taylor series expansion of $\eta(x^{(i)}; \cdot)$, we have

$$\begin{aligned} q^* \;&\geq \\ &\Lambda_{min}\left[\frac{1}{n}\sum_{i=1}^n \eta(x^{(i)}; \theta_S^*) x_S^{(i)} (x_S^{(i)})^T\right] - \max_{\alpha \in [0,1]} \Big\|\frac{1}{n}\sum_{i=1}^n \eta'(x^{(i)}; \theta_S^* + \alpha u_S)(u_S^T x_S^{(i)}) x_S^{(i)} (x_S^{(i)})^T\Big\|_2 \\ =\;& \Lambda_{min}(Q_{SS}^*) - \max_{\alpha \in [0,1]} \Big\|\frac{1}{n}\sum_{i=1}^n \eta'(x^{(i)}; \theta_S^* + \alpha u_S)(u_S^T x_S^{(i)}) x_S^{(i)} (x_S^{(i)})^T\Big\|_2 \\ \geq\;& C_{min} - \max_{\alpha \in [0,1]} \Big\|\frac{1}{n}\sum_{i=1}^n \eta'(x^{(i)}; \theta_S^* + \alpha u_S)(u_S^T x_S^{(i)}) x_S^{(i)} (x_S^{(i)})^T\Big\|_2 \end{aligned}$$

It remains to control the final spectral norm. For any fixed $\alpha \in [0,1]$ and $y \in \mathbb{R}^d$ with $\|y\|_2 = 1$, we have

$$\begin{aligned} y^T \left\{\frac{1}{n}\sum_{i=1}^n \eta'(x^{(i)}; \theta_S^* + \alpha u_S)(u_S^T x_S^{(i)}) x_S^{(i)} (x_S^{(i)})^T\right\} y \;&= \\ \frac{1}{n}\sum_{i=1}^n \eta'(x^{(i)}; \theta_S^* + \alpha u_S)(u_S^T x_S^{(i)}) \;&\left[x_S^{(i)})^T y\right]^2 \\ \leq\; \frac{1}{n}\sum_{i=1}^n \left|\eta'(x^{(i)}; \theta_S^* + \alpha u_S)(u_S^T x_S^{(i)})\right| \;&\left[x_S^{(i)})^T y\right]^2 \end{aligned}$$

Now note that $|\eta'(x^{(i)}; \theta_S^* + \alpha u_S)| \leq 1$, and $|u_S^T x_S^{(i)}| \leq \sqrt{d}\|u_S\|_2 = M\lambda_n d$. Moreover, we have $\|\frac{1}{n}\sum_{i=1}^n \left(x_S^{(i)})^T y\right) \leq \|\frac{1}{n}\sum_{i=1}^n x_S^{(i)} (x_S^{(i)})^T\|_2 \;\leq\; D_{\max}$ by assumption. Combining these pieces, we obtain

$$\begin{aligned} \max_{\alpha \in [0,1]} \Big\|\frac{1}{n}\sum_{i=1}^n \eta'(x^{(i)}; \theta_S^* + \alpha u_S)(u_S^T x_S^{(i)}) x_S^{(i)} (x_S^{(i)})^T\Big\|_2 \;&\leq\; D_{\max} M \lambda_n d \\ &\leq\; \frac{C_{min}}{2}, \end{aligned}$$

where the last inequality follows as long as $\lambda_n d \leq \frac{C_{min}}{2D_{\max}M}$. We have thus shown that

$$q^* := \Lambda_{min}(\nabla^2 \ell(\theta^*_S + \alpha u; \{x^{(i)}\})) \quad \geq \quad \frac{C_{min}}{2} \tag{53}$$

with probability converging to one, as long as $\lambda_n d$ is sufficiently small.

Finally, combining the bounds (51), (52), and (53) in the expression (50), we conclude that

$$G(u_S) \quad \geq \quad \left(\lambda_n \sqrt{d}\right)^2 \left\{ -\frac{1}{4}M + \frac{C_{min}}{2}M^2 - M \right\}.$$

This expression is strictly positive for $M = 5/C_{min}$. Moreover, for this choice of M, we have that $\lambda_n d$ must be upper bounded by $\frac{C_{min}}{2D_{\max}M} = \frac{C^2_{min}}{10D_{\max}}$, as assumed in the lemma statement.

A.3 Proof of Lemma 4

We first show that the remainder term R^n satisfies the bound $\|R^n\|_\infty \leq D_{\max}\|\widehat{\theta}_S - \theta^*_S\|_2^2$. Then the result of Lemma 3—namely, that $\|\widehat{\theta}_S - \theta^*_S\|_2 = \mathcal{O}_p(\lambda_n\sqrt{d})$—can be used to conclude that $\frac{\|R^n\|_\infty}{\lambda_n} = \mathcal{O}_p(\lambda_n d)$, which suffices to guarantee the claim of Lemma 4.

Focusing on element R^n_j for some index $j \in \{1, \ldots, p\}$, we have

$$\begin{aligned} R^n_j &= \left[\nabla^2 \ell(\bar{\theta}^{(j)}; x) - \nabla^2 \ell(\theta^*; x)\right]^T_j [\widehat{\theta} - \theta^*] \\ &= \frac{1}{n}\sum_{i=1}^n \left[\eta(x^{(i)}; \bar{\theta}^{(j)}) - \eta(x^{(i)}; \theta^*)\right] \left[x^{(i)}(x^{(i)})^T\right]^T_j [\widehat{\theta} - \theta^*]. \end{aligned}$$

for some point $\bar{\theta}^{(j)} = t_j\widehat{\theta} + (1 - t_j)\theta^*$. Setting $g(t) = \frac{4\exp(2t)}{[1+\exp(2t)]^2}$, note that $\eta(x;\theta) = g(x_r \sum_{t \in V \setminus r} \theta_{rt} x_t)$. By the chain rule and another application of the mean value theorem, we then have

$$\begin{aligned} R^n_j &= \frac{1}{n}\sum_{i=1}^n g'(\bar{\bar{\theta}}^{(j)T} x^{(i)})(x^{(i)})^T[\bar{\theta}^{(j)} - \theta^*] \left\{ x^{(i)}_j (x^{(i)})^T [\widehat{\theta} - \theta^*] \right\} \\ &= \frac{1}{n}\sum_{i=1}^n \left\{ g'(\bar{\bar{\theta}}^{(j)T} x^{(i)}) x^{(i)}_j \right\} \left\{ \bar{\theta}^{(j)} - \theta^*]^T x^{(i)} (x^{(i)})^T [\widehat{\theta} - \theta^*] \right\} \end{aligned}$$

where $\bar{\bar{\theta}}^{(j)}$ is another point on the line joining $\widehat{\theta}$ and θ^*. Setting $a_i := \{g'(\bar{\bar{\theta}}^{(j)T} x^{(i)}) x^{(i)}_j\}$ and $b_i := \{[\bar{\theta}^{(j)} - \theta^*]^T x^{(i)} (x^{(i)})^T [\widehat{\theta} - \theta^*]\}^2$, we have

$$|R^n_j| \quad = \quad \frac{1}{n}\left|\sum_{i=1}^n a_i b_i\right| \quad \leq \quad \frac{1}{n}\|a\|_\infty \|b\|_1.$$

A calculation shows that $\|a\|_\infty \leq 1$, and

$$\begin{aligned}
\frac{1}{n}\|b\|_1 &= t_j[\widehat{\theta} - \theta^*]^T \left\{ \frac{1}{n}\sum_{i=1}^n x^{(i)}(x^{(i)})^T \right\} [\widehat{\theta} - \theta^*] \\
&= t_j[\widehat{\theta}_S - \theta^*_S]^T \left\{ \frac{1}{n}\sum_{i=1}^n x_S^{(i)}(x_S^{(i)})^T \right\} [\widehat{\theta}_S - \theta^*_S] \\
&\leq D_{\max}\|\widehat{\theta}_S - \theta^*_S\|_2^2,
\end{aligned}$$

where the second line uses the fact that $\widehat{\theta}_{S^c} = \theta^*_{S^c} = 0$. This concludes the proof.

B Proof of Lemma 7

Recall from the discussion leading up to the bound (42) that element (j,k) of the matrix difference $Q^n - Q^*$, denoted by Z_{jk}, satisfies a sharp tail bound. By definition of the ℓ_∞-matrix norm, we have

$$\begin{aligned}
\mathbb{P}[|||Q^n_{S^cS} - Q^*_{S^cS}|||_\infty \geq \delta] &= \mathbb{P}[\max_{j \in S^c} \sum_{k \in S} |Z_{jk}| \geq \delta] \\
&\leq (p-d)\, \mathbb{P}[\sum_{k \in S} |Z_{jk}| \geq \delta],
\end{aligned}$$

where the final inequality uses a union bound, and the fact that $|S^c| \leq p - d$. Via another union bound over the row elements, we have

$$\mathbb{P}[|||Q^n_{S^cS} - Q^*_{S^cS}|||_\infty \geq \delta] \quad \leq (p-d)\; d\; \mathbb{P}\left[|Z_{jk}| \geq \delta/d\right],$$

from which the claim (45a) follows by setting $\epsilon = \delta/d$ in the Hoeffding bound (42). The proof of bound (45b) is analogous, with the pre-factor $(p-d)$ replaced by d.

To prove the last claim (45c), we write

$$\begin{aligned}
|||(Q^n_{SS})^{-1} - (Q^*_{SS})^{-1}|||_\infty &= |||(Q^*_{SS})^{-1}\left[Q^*_{SS} - Q^n_{SS}\right](Q^n_{SS})^{-1}|||_\infty \\
&\leq \sqrt{d}\; |||(Q^*_{SS})^{-1}\left[Q^*_{SS} - Q^n_{SS}\right](Q^n_{SS})^{-1}|||_2 \\
&\leq \sqrt{d}\; |||(Q^*_{SS})^{-1}|||_2 |||Q^*_{SS} - Q^n_{SS}|||_2 |||(Q^n_{SS})^{-1}|||_2 \\
&\leq \frac{\sqrt{d}}{C_{min}} |||Q^*_{SS} - Q^n_{SS}|||_2 |||(Q^n_{SS})^{-1}|||_2.
\end{aligned}$$

From the proof of Lemma 5, in particular equation (43), we have

$$\mathbb{P}\left[|||(Q^n_{SS})^{-1}|||_2 \geq \frac{2}{C_{min}}\right] \quad \leq 2\exp\left(-K\tfrac{\delta^2 n}{d^2} + B\log(d)\right)$$

for a constants B. Moreover, from equation (43), we have

$$\mathbb{P}[|||Q^n_{SS} - Q_{SS}|||_2 \geq \delta/\sqrt{d}] \quad \leq \quad 2\exp\left(-K\frac{\delta^2 n}{d^3} + 2\log(d)\right),$$

so that the bound (45c) follows.

References

P. Abbeel, D. Koller, and A. Y. Ng. Learning factor graphs in polynomial time and sample complexity. *Jour. Mach. Learning Res.*, 7:1743–1788, 2006.

D.P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, 1995.

G. Bresler, E. Mossel, and A. Sly. Reconstruction of Markov random fields from samples: Some easy observations and algorithms. `http://front.math.ucdavis.edu/0712.1402`, 2008.

E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 2006.

D. Chickering. Learning Bayesian networks is NP-complete. *Proceedings of AI and Statistics*, 1995.

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

I. Csiszár and Z. Talata. Consistent estimation of the basic neighborhood structure of Markov random fields. *The Annals of Statistics*, 34(1):123–145, 2006.

S. Dasgupta. Learning polytrees. In *Uncertainty on Artificial Intelligence*, pages 134–14, 1999.

K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–336. Elsevier, Amsterdan, NL, 2001.

D. Donoho and M. Elad. Maximal sparsity representation via ℓ_1 minimization. *Proc. Natl. Acad. Sci.*, 100:2197–2202, March 2003.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

K. Koh, S. J. Kim, and S. Boyd. An interior-point method for large-scale ℓ_1-regularized logistic regression. *Jour. Mach. Learning Res.*, 3:1519–1555, 2007.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3), 2006.

A. Y. Ng. Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance. In *International Conference on Machine Learning*, 2004.

G. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal on Statistics*, 2008. To appear.

N. P. Santhanam and M. J. Wainwright. Information-theoretic limits of high-dimensional model selection. In *International Symposium on Information Theory*, Toronto, Canada, July 2008.

P. Spirtes, C. Glymour, and R. Scheines. Causation, prediction and search. *MIT Press*, 2000.

N. Srebro. Maximum likelihood bounded tree-width Markov networks. *Artificial Intelligence*, 143(1):123–138, 2003.

J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE Trans. Info. Theory*, 51(3):1030–1051, March 2006.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programs. In *Proc. Allerton Conference on Communication, Control and Computing*, October 2006. Long version: UC Berkeley Technical Report 709, May 2006.

M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1465–1472. MIT Press, Cambridge, MA, 2007.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. of Mach. Learn. Res.*, 7: 2541–2567, 2007.