The Annals of Statistics
2010, Vol. 38, No. 3, 1287–1319
DOI: 10.1214/09-AOS691

HIGH-DIMENSIONAL ISING MODEL SELECTION USING ℓ_1-REGULARIZED LOGISTIC REGRESSION

By Pradeep Ravikumar[1,2,3], Martin J. Wainwright[3] and John D. Lafferty[1]

University of California, Berkeley, University of California, Berkeley and Carnegie Mellon University

We consider the problem of estimating the graph associated with a binary Ising 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. The method is analyzed under high-dimensional scaling in which both the number of nodes p and maximum neighborhood size 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) and the model parameters for the method to succeed in consistently estimating the neighborhood of every node in the graph simultaneously. With coherence conditions imposed 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 exponentially decaying error. When these same conditions are imposed directly on the sample matrices, we show that a reduced sample size of $n = \Omega(d^2 \log p)$ suffices for the method to estimate neighborhoods consistently. Although this paper focuses on the binary graphical models, we indicate how a generalization of the method of the paper would apply to general discrete Markov random fields.

1. Introduction. Undirected graphical models, also known as Markov random fields, are used in a variety of domains, including statistical physics [17], natural language processing [21], image analysis [8, 14, 37] and spatial

Received October 2008; revised January 2009.
[1]Supported in part by NSF Grants IIS-0427206 and CCF-0625879.
[2]Supported in part by a Siebel Scholarship.
[3]Supported in part by NSF Grants DMS-06-05165 and CCF-0545862.
AMS 2000 subject classifications. Primary 62F12; secondary 68T99.
Key words and phrases. Graphical models, Markov random fields, structure learning, ℓ_1-regularization, model selection, convex risk minimization, high-dimensional asymptotics.

statistics [26], 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 $i \in V$. One important problem for such models is to estimate the underlying graph from n independent and identically distributed samples $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ drawn from the distribution specified by some Markov random field. As a concrete illustration, for binary random variables, each vector-valued sample $x^{(i)} \in \{0, 1\}^p$ might correspond to the votes of a set of p politicians on a particular bill, and estimating the graph structure amounts to detecting statistical dependencies in these voting patterns (see Banerjee, Ghaoui and d'Asprémont [2] for further discussion of this example).

Due to both its importance and difficulty, the problem of structure learning for discrete graphical models has attracted considerable attention. The absence of an edge in a graphical model encodes a conditional independence assumption. Constraint-based approaches [30] 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 measure of the goodness of fit 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 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 [6] shows that this problem is in general NP-hard.

A complication for undirected graphical models involving discrete random variables is that typical score metrics involve the partition function or cumulant function associated with the Markov random field. For general undirected MRFs, calculation of this partition function is computationally intractable [36]. The space of candidate structures in scoring based approaches is thus typically restricted to either directed graphical models [10] or to simple sub-classes of undirected graphical models such as those based on trees [7] and hypertrees [31]. Abbeel, Koller and Ng [1] 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 number of vertices 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 [9] show consistency of a

method that uses pseudo-likelihood and a modification of the BIC criterion, but this also involves a prohibitively expensive search.

The main contribution of this paper is a careful analysis of the computational and statistical efficiency of a simple method for graphical model selection. 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. 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 on the population Fisher information matrix, consistent neighborhood selection is possible using $n = \Omega(d^3 \log p)$ samples and computational complexity $\mathcal{O}(\max\{n, p\}p^3)$. We also show that when the same assumptions are imposed directly on the sample matrices, $n = \Omega(d^2 \log p)$ samples suffice for consistent neighborhood selection with the same computational complexity. We focus in this paper on binary Ising models, but indicate in Section 7 a generalization of the method applicable to general discrete Markov random fields.

The technique of ℓ_1-regularization for estimation of sparse models or signals has a long history in many fields (for instance, see [32] 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., [5, 12, 23, 24, 32, 33, 39]). Despite the well-known computational intractability of computing marginals and likelihoods for discrete MRFs [36], 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 $\mathcal{O}(\max\{p, n\}p^3)$ and is thus well suited to high-dimensional problems [20]. Conceptually, like the work of Meinshausen and Bühlmann [23] 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 [36].

Portions of this work were initially reported in a conference publication [35], with the weaker result that $n = \Omega(d^6 \log d + d^5 \log p)$ samples suffice for consistent Ising model selection. Since the appearance of that paper, other researchers have also studied the problem of model selection in discrete Markov random fields. For the special case of bounded degree models, Bresler, Mossel and Sly [4] describe a simple search-based method, and prove under relatively mild assumptions that it can recover the graph structure

with $\Theta(\log p)$ samples. However, in the absence of additional restrictions, the computational complexity of the method is $\mathcal{O}(p^{d+1})$. In other work, Santhanam and Wainwright [29] 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 costs.

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 whereas Section 5 provides concentration results linking the population matrices to the sample versions. In Section 6, we provide some experimental results that illustrate the practical performance of our method and the close agreement between theory and practice. Section 7 discusses an extension to more general Markov random fields, and we conclude in Section 8.

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 $v \in \mathbb{R}^d$ and parameter $q \in [1, \infty]$, we use $\|v\|_q$ to denote the usual ℓ_q norm. Given a matrix $A \in \mathbb{R}^{a \times b}$ and parameter $q \in [1, \infty]$, we use $|\!|\!| A |\!|\!|_q$ to denote the induced matrix-operator norm with A viewed as a mapping from $\ell_q^b \to \ell_q^a$ (see Horn and Johnson [16]). Two examples of particular importance in this paper are the spectral norm $|\!|\!| A |\!|\!|_2$, corresponding to the maximal singular value of A, and the ℓ_∞ matrix norm, given by $|\!|\!| A |\!|\!|_\infty = \max_{j=1,\ldots,a} \sum_{k=1}^b |A_{jk}|$. We make use of the bound $|\!|\!| A |\!|\!|_\infty \leq \sqrt{a} |\!|\!| A |\!|\!|_2$ for any symmetric matrix $A \in \mathbb{R}^{a \times a}$.

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. *Pairwise Markov random fields.* Let $X = (X_1, X_2, \ldots, X_p)$ denote a random vector with each variable X_s taking values in a corresponding set $\mathcal{X}_s$. Say we are given an undirected graph G with vertex set $V = \{1, \ldots, p\}$ and edge set E, so that each random variable X_s is associated with a vertex $s \in V$. The pairwise Markov random field associated with the graph G over the random vector X is the family of distributions of X which factorize as $\mathbb{P}(x) \propto$

$\exp\{\sum_{(s,t)\in E}\phi_{st}(x_s,x_t)\}$ where for each edge $(s,t)\in E$, ϕ_{st} is a mapping from pairs $(x_s,x_t)\in\mathcal{X}_s\times\mathcal{X}_t$ to the real line. For models involving discrete random variables, the pairwise assumption involves no loss of generality since any Markov random field with higher-order interactions can be converted (by introducing additional variables) to an equivalent Markov random field with purely pairwise interactions (see Wainwright and Jordan [34] for details of this procedure).

Ising model. In this paper, we focus on the special case of the Ising model in which $X_s\in\{-1,1\}$ for each vertex $s\in V$, and $\phi_{st}(x_s,x_t)=\theta^*_{st}x_sx_t$ for some parameter $\theta^*_{st}\in\mathbb{R}$, so that the distribution takes the form

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

The partition function $Z(\theta^*)$ ensures that the distribution sums to one. This model is used in many applications of spatial statistics such as modeling the behavior of gases or magnets in statistical physics [17], building statistical models in computer vision [13] and social network analysis.

2.2. *Graphical model selection.* Suppose that we are given a collection $\mathfrak{X}_1^n:=\{x^{(1)},\ldots,x^{(n)}\}$ of n samples where each p-dimensional vector $x^{(i)}\in\{-1,+1\}^p$ is drawn in an i.i.d. manner from a distribution $\mathbb{P}_{\theta^*}$ of the form (1) for parameter vector θ^* and graph $G=(V,E)$ over the p variables. It is convenient to view the parameter vector θ^* as a $\binom{p}{2}$-dimensional vector, indexed by pairs of distinct vertices but nonzero 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. 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

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

Here the sign function takes value $+1$ if $\theta^*_{st}>0$, value -1 if $\theta^*_{st}<0$ and 0, otherwise. 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—among them gene microarray data and social network analysis—the model dimension p is comparable to or larger than the sample size n, so that the relevance of such "fixed p" asymptotics is limited. With this motivation, our analysis in this paper is of the high-dimensional nature,

in which both the model dimension and the sample size are allowed to increase, and we study the scalings 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

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

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.* 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 $\operatorname{sign}(\theta^*_{rt})$ for all $t \in \mathcal{N}(r)$. To capture both the neighborhood structure and sign pattern, we define the product set of "signed vertices" as $\{-1, 1\} \times V$. We use the shorthand "ιr" for elements $(\iota, r) \in \{-1, 1\} \times V$. We then define the *signed neighborhood set* as

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

Here the sign function has an unambiguous definition, since $\theta^*_{rt} \neq 0$ for all $t \in \mathcal{N}(r)$. 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

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

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 $X_{\backslash r} = \{X_t \mid t \in V \setminus \{r\}\}$. A simple calculation shows that under the model (1), this conditional distribution takes the form

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

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 an ℓ_1-regularized logistic regression of X_r on the other variables $X_{\backslash r}$. Explicitly, given $\mathfrak{X}_1^n = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, a set of n i.i.d. samples, this regularized regression problem is a convex program of the form

$$\min_{\theta_{\backslash r} \in \mathbb{R}^{p-1}} \{\ell(\theta; \mathfrak{X}_1^n) + \lambda_{(n,p,d)} \|\theta_{\backslash r}\|_1\}, \tag{5}$$

where

$$\ell(\theta; \mathfrak{X}_1^n) := -\frac{1}{n} \sum_{i=1}^n \log \mathbb{P}_\theta(x_r^{(i)} \mid x_{\backslash r}^{(i)}) \tag{6}$$

is the rescaled negative log likelihood (the rescaling factor $1/n$ in this definition is for later theoretical convenience) and $\lambda_{(n,p,d)} > 0$ is a regularization parameter, to be specified by the user. For notational convenience, we will also use λ_n as notation for this regularization parameter suppressing the potential dependence on p and d.

Following some algebraic manipulation, the regularized negative log likelihood can be written as

$$\min_{\theta_{\backslash r} \in \mathbb{R}^{p-1}} \left\{ \frac{1}{n} \sum_{i=1}^n f(\theta; x^{(i)}) - \sum_{u \in V \backslash r} \theta_{ru} \widehat{\mu}_{ru} + \lambda_n \|\theta_{\backslash r}\|_1 \right\}, \tag{7}$$

where

$$f(\theta; x) := \log\left\{ \exp\left(\sum_{t \in V \backslash r} \theta_{rt} x_t \right) + \exp\left(- \sum_{t \in V \backslash r} \theta_{rt} x_t \right) \right\} \tag{8}$$

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 (7) is convex but not differentiable, due to the presence of the ℓ_1-regularizer. By Lagrangian duality, the problem (7) 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 $\widehat{\theta}^n_{\backslash r}$ be an element of the minimizing set of problem (7). Although $\widehat{\theta}^n_{\backslash r}$ need not be unique in general since the problem (7) need not be strictly convex, our analysis shows that in the regime of interest, this minimizer $\widehat{\theta}^n_{\backslash r}$ is indeed unique. We use $\widehat{\theta}^n_{\backslash r}$ to estimate the signed neighborhood $\mathcal{N}_\pm(r)$ according to

$$\widehat{\mathcal{N}}_\pm(r) := \{\operatorname{sign}(\widehat{\theta}_{ru}) u \mid u \in V \setminus r, \widehat{\theta}_{su} \neq 0\}. \tag{9}$$

We say that the full graph G is estimated consistently, written as the event $\{\widehat{E}_n = E^*\}$, if every signed neighborhood is recovered—that is, $\widehat{\mathcal{N}}_\pm(r) = \mathcal{N}_\pm(r)$ 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 (9), one for each node r of the graph, agree with the true neighborhoods so that the full graph is estimated consistently. In this section, we begin by stating the assumptions that underlie our analysis, 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 details 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 sufficiently fast rates, we can then use a union bound over all p nodes of the graph to conclude that consistent graph selection 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

$$Q^*_r := \mathbb{E}_{\theta^*}\{\nabla^2 \log \mathbb{P}_{\theta^*}[X_r \mid X_{\setminus r}]\}. \tag{10}$$

For future reference, this is given as the explicit expression

$$Q^*_r = \mathbb{E}_{\theta^*}[\eta(X;\theta^*) X_{\setminus r} X^T_{\setminus r}], \tag{11}$$

where

$$\eta(u;\theta) := \frac{4\exp(2u_r \sum_{t \in V\setminus r} \theta_{rt} u_t)}{(\exp(2u_r \sum_{t \in V\setminus r} \theta_{rt} u_t) + 1)^2} \tag{12}$$

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 [23, 32, 39].

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; that is, there exists a constant $C_{\min} > 0$ such that

$$\Lambda_{\min}(Q^*_{SS}) \geq C_{\min}. \tag{13}$$

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

(A2) Incoherence condition. Our next assumption captures the intuition that the large number of irrelevant covariates (i.e., nonneighbors 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

$$|||Q^*_{S^c S}(Q^*_{SS})^{-1}|||_\infty \leq 1 - \alpha. \tag{14}$$

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

$$\theta^*_{\min} = \min_{(r,t)\in E} |\theta^*_{rt}|. \tag{15}$$

With this definition, we have the following:

THEOREM 1. *Consider an Ising graphical model with parameter vector θ^* and associated edge set E^* such that conditions* (A1) *and* (A2) *are satisfied by the population Fisher information matrix Q^*, and let $\mathfrak{X}^n_1$ be a set of n i.i.d. samples from the model specified by θ^*. Suppose that the regularization parameter λ_n is selected to satisfy*

$$\lambda_n \geq \frac{16(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}. \tag{16}$$

Then there exist positive constants L and K, independent of (n, p, d), such that if

$$n > Ld^3 \log p, \tag{17}$$

then the following properties hold with probability at least $1 - 2\exp(-K\lambda_n^2 n)$.

(a) *For each node $r \in V$, the ℓ_1-regularized logistic regression (5), given data $\mathfrak{X}_1^n$, has a unique solution, and so uniquely specifies a signed neighborhood $\widehat{\mathcal{N}}_\pm(r)$.*

(b) *For each $r \in V$, the estimated signed neighborhood $\widehat{\mathcal{N}}_\pm(r)$ correctly excludes all edges not in the true neighborhood. Moreover, it correctly includes all edges (r,t) for which $|\theta^*_{rt}| \geq \frac{10}{C_{\min}}\sqrt{d}\lambda_n$.*

The theorem not only specifies sufficient conditions but also the probability with which the method recovers the true signed edge-set. This probability decays exponentially as a function of $\lambda_n^2 n$ which leads naturally to the following corollary on model selection consistency of the method for a sequence of Ising models specified by $(n, p(n), d(n))$.

COROLLARY 1. *Consider a sequence of Ising models with graph edge sets $\{E^*_{p(n)}\}$ and parameters $\{\theta^*_{(n,p,d)}\}$; each of which satisfies conditions* (A1) *and* (A2). *For each n, let $\mathfrak{X}_1^n$ be a set of n i.i.d. samples from the model specified by $\theta^*_{(n,p,d)}$, and suppose that $(n, p(n), d(n))$ satisfies the scaling condition (17) of Theorem 1. Suppose further that the sequence $\{\lambda_n\}$ of regularization parameters satisfies condition (16) and*

$$\lambda_n^2 n \to \infty \tag{18}$$

and the minimum parameter weights satisfy

$$\min_{(r,t)\in E^*_n} |\theta^*_{(n,p,d)}(r,t)| \geq \frac{10}{C_{\min}}\sqrt{d}\lambda_n \tag{19}$$

for sufficiently large n. Then the method is model selection consistent so that if $\widehat{E}_{p(n)}$ is the graph structure estimated by the method given data $\mathfrak{X}_1^n$, then $\mathbb{P}[\widehat{E}_{p(n)} = E^*_{p(n)}] \to 1$ as $n \to \infty$.

Remarks. (a) It is worth noting that the scaling condition (17) on (n,p,d) allows for graphs and sample sizes in the "large p, small n" regime (meaning $p \gg n$), as long as the degrees are bounded, or grow at a sufficiently slow rate. In particular, one set of sufficient conditions are the scalings

$$d = O(n^{c_1}) \quad \text{and} \quad p = O(e^{n^{c_2}}), \qquad 3c_1 + c_2 < 1,$$

for some constants $c_1, c_2 > 0$. Under these scalings, note that we have $d^3 \log(p) = \mathcal{O}(n^{3c_1+c_2}) = o(n)$, so that condition (17) holds.

A bit more generally, note that in the regime $p \gg n$, the growth condition (17) requires that that $d = o(p)$. However, in many practical applications of graphical models (e.g., image analysis, social networks), one is interested in node degrees d that remain bounded or grow sub-linearly in the graph size so that this condition is not unreasonable.

(b) Loosely stated, the theorem requires that the edge weights are not too close to zero (in absolute value) for the method to estimate the true graph. In particular, conditions (16) and (19) imply that the minimum edge weight $\theta^*_{\min}$ is required to scale as

$$\theta^*_{\min} = \Omega\left(\sqrt{\frac{d\log p}{n}}\right).$$

Note that in the classical fixed (p, d) case, this reduces to the familiar scaling requirement of $\theta^*_{\min} = \Omega(n^{-1/2})$.

(c) In the high-dimensional setting (for $p \to +\infty$), a choice of the regularization parameter satisfying both conditions (16) and (18) is, for example,

$$\lambda_n = \frac{16(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$$

for which the probability of incorrect model selection decays at rate $\mathcal{O}(\exp(-K'\log p))$ for some constant $K' > 0$. In the classical setting (fixed p), this choice can be modified to $\lambda_n = \frac{16(2-\alpha)}{\alpha}\sqrt{\frac{\log(pn)}{n}}$.

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 condition (A1) and the mutual incoherence condition (A2) hold for the *sample Fisher information matrix*

$$Q^n := \widehat{\mathbb{E}}[\eta(X;\theta^*)X_{\backslash r}X_{\backslash r}^T] = \frac{1}{n}\sum_{i=1}^{n}\eta(x^{(i)};\theta^*)x_{\backslash r}^{(i)}(x_{\backslash r}^{(i)})^T, \tag{20}$$

then the growth condition (17) and choice of λ_n from Theorem 1 are sufficient to ensure that the graph is recovered with high probability.

The second part of the analysis, provided in Section 5, is devoted to showing that under the specified growth condition (17), 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. On one hand, 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. However, in the current setting, the added delicacy is that we are required to control this convergence over subsets of increasing size. Our proof therefore requires some large-deviation analysis for random matrices with dependent elements so as to provide exponential control on the rates of convergence.

3.3. *Primal-dual witness for graph recovery.* At the core of our proof lies the notion of a primal-dual witness used in previous work on the Lasso [33]. In particular, our proof involves the explicit construction of an optimal *primal-dual pair*—namely, a primal solution $\widehat{\theta} \in \mathbb{R}^{p-1}$ along with an associated subgradient vector $\widehat{z} \in \mathbb{R}^{p-1}$ (which can be interpreted as a dual solution), such that the sub-gradient optimality conditions associated with the convex program (7) 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.

For the convex program (7), the zero sub-gradient optimality conditions [27] take the form

$$\nabla \ell(\widehat{\theta}) + \lambda_n \widehat{z} = 0, \tag{21}$$

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

$$\widehat{z}_{rt} = \operatorname{sign}(\widehat{\theta}_{rt}) \qquad \text{if } \widehat{\theta}_i \neq 0 \quad \text{and} \quad |\widehat{z}_{rt}| \leq 1 \qquad \text{otherwise.} \tag{22}$$

By convexity, a pair $(\widehat{\theta}, \widehat{z}) \in \mathbb{R}^{p-1} \times \mathbb{R}^{p-1}$ is a primal-dual optimal solution to the convex program and its dual if and only if the two conditions (21) and (22) are satisfied. Of primary interest to us is the property that such an optimal primal-dual pair correctly specifies the signed neighborhood of node r; the necessary and sufficient conditions for such correctness are

$$\operatorname{sign}(\widehat{z}_{rt}) = \operatorname{sign}(\theta^*_{rt}) \qquad \forall (r,t) \in S := \{(r,t) \in E\} \quad \text{and} \tag{23a}$$

$$\widehat{\theta}_{ru} = 0 \qquad \text{for all } (r,u) \in S^c := E \setminus S. \tag{23b}$$

The ℓ_1-regularized logistic regression problem (7) is convex; however, for $p \gg n$, it need not be strictly convex, so that there may be multiple optimal solutions. The following lemma, proved in Appendix A, provides sufficient conditions for shared sparsity among optimal solutions, as well as uniqueness of the optimal solution:

LEMMA 1. *Suppose that 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 $[\nabla^2 \ell(\widehat{\theta})]_{SS}$ is strictly positive definite, then $\widehat{\theta}$ is the unique optimal solution.*

Based on this lemma, we construct a primal-dual witness $(\widehat{\theta}, \widehat{z})$ with the following steps.

(a) First, we set $\widehat{\theta}_S$ as the minimizer of the partial penalized likelihood

$$\widehat{\theta}_S = \operatorname*{arg\,min}_{(\theta_S, 0) \in \mathbb{R}^{p-1}} \{\ell(\theta; \mathfrak{X}_1^n) + \lambda_n \|\theta_S\|_1\} \tag{24}$$

and set $\widehat{z}_S = \operatorname{sign}(\widehat{\theta}_S)$.

(b) Second, we set $\widehat{\theta}_{S^c} = 0$ so that condition (23b) holds.
(c) In the third step, we obtain $\widehat{z}_{S^c}$ from (21) by substituting in the values of $\widehat{\theta}$ and $\widehat{z}_S$. Thus our construction satisfies conditions (23b) and (21).
(d) The final and most challenging step consists of showing that the stated scalings of (n, p, d) imply that, with high-probability, the remaining conditions (23a) and (22) are satisfied.

Our analysis in step (d) guarantees that $\|\widehat{z}_{S^c}\|_\infty < 1$ with high probability. 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 by Lemma 1, the primal solution $\widehat{\theta}$ is guaranteed to be unique.

It should be noted that, since S is unknown, the primal-dual witness method is *not* a practical algorithm that could ever be implemented to solve ℓ_1-regularized logistic regression. Rather, it is a proof technique that allows us to establish sign correctness of the unique optimal solution.

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, recalling the definition (20) of the sample Fisher information matrix $Q^n = \widehat{\mathbb{E}}[\nabla^2 \ell(\theta^*)]$, we define the "good event,"

$$\mathcal{M}(\mathfrak{X}_1^n) := \{\mathfrak{X}_1^n \in \{-1, +1\}^{n \times p} \mid Q^n \text{ satisfies (A1) and (A2)}\}. \tag{25}$$

As in the statement of Theorem 1, the quantities L and K refer to constants independent of (n, p, d). With this notation, we have the following:

PROPOSITION 1 (Fixed design). *If the event $\mathcal{M}(\mathfrak{X}_1^n)$ holds, the sample size satisfies $n > Ld^2 \log(p)$, and the regularization parameter is chosen such that $\lambda_n \geq \frac{16(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$, then with probability at least $1 - 2\exp(-K\lambda_n^2 n) \to 1$, the following properties hold.*

(a) *For each node $r \in V$, the ℓ_1-regularized logistic regression has a unique solution, and so uniquely specifies a signed neighborhood $\widehat{\mathcal{N}}_\pm(r)$.*

(b) *For each $r \in V$, the estimated signed neighborhood vector $\widehat{\mathcal{N}}_\pm(r)$ correctly excludes all edges not in the true neighborhood. Moreover, it correctly includes all edges with $|\theta_{rt}| \geq \frac{10}{C_{\min}}\sqrt{d}\lambda_n$.*

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 B. The central object is the following expansion obtained by re-writing the zero-subgradient condition as

$$\nabla \ell(\widehat{\theta}; \mathfrak{X}_1^n) - \nabla \ell(\theta^*; \mathfrak{X}_1^n) = W^n - \lambda_n \widehat{z}, \tag{26}$$

where we have introduced the short-hand notation $W^n = -\nabla \ell(\theta^*; \mathfrak{X}_1^n)$ for the $(p-1)$-dimensional score function,

$$W^n := -\frac{1}{n} \sum_{i=1}^n x_{\backslash r}^{(i)} \left\{ x_r^{(i)} - \frac{\exp(\sum_{t \in V \backslash r} \theta_{rt}^* x_t^{(i)}) - \exp(-\sum_{t \in V \backslash r} \theta_{rt}^* x_t^{(i)})}{\exp(\sum_{t \in V \backslash r} \theta_{rt}^* x_t^{(i)}) + \exp(-\sum_{t \in V \backslash r} \theta_{rt}^* x_t^{(i)})} \right\}.$$

For future reference, note that $\mathbb{E}_{\theta^*}[W^n] = 0$. Next, applying the mean-value theorem coordinate-wise to the expansion (26) yields

$$\nabla^2 \ell(\theta^*; \mathfrak{X}_1^n)[\widehat{\theta} - \theta^*] = W^n - \lambda_n \widehat{z} + R^n, \tag{27}$$

where the remainder term takes the form

$$R_j^n = [\nabla^2 \ell(\bar{\theta}^{(j)}; \mathfrak{X}_1^n) - \nabla^2 \ell(\theta^*; \mathfrak{X}_1^n)]_j^T (\widehat{\theta} - \theta^*) \tag{28}$$

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

LEMMA 2. *For the specified mutual incoherence parameter $\alpha \in (0, 1]$, we have*

$$\mathbb{P}\left(\frac{2-\alpha}{\lambda_n} \|W^n\|_\infty \geq \frac{\alpha}{4} \right) \leq 2 \exp\left(-\frac{\alpha^2 \lambda_n^2}{128(2-\alpha)^2} n + \log(p) \right), \tag{29}$$

which converges to zero at rate $\exp(-c\lambda_n^2 n)$ as long as $\lambda_n \geq \frac{16(2-\alpha)}{\alpha} \sqrt{\frac{\log p}{n}}$.

See Appendix B.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 $\lambda_n d \leq \frac{C_{\min}^2}{10 D_{\max}}$ and $\|W^n\|_\infty \leq \lambda_n/4$, then*

$$\|\widehat{\theta}_S - \theta_S\|_2 \leq \frac{5}{C_{\min}} \sqrt{d} \lambda_n. \tag{30}$$

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

Our final technical lemma provides control on the remainder term (28).

LEMMA 4. *If* $\lambda_n d \leq \frac{C_{\min}^2}{100 D_{\max}} \frac{\alpha}{2-\alpha}$ *and* $\|W^n\|_\infty \leq \lambda_n/4$, *then*

$$\frac{\|R^n\|_\infty}{\lambda_n} \leq \frac{25 D_{\max}}{C_{\min}^2} \lambda_n d \leq \frac{\alpha}{4(2-\alpha)}.$$

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

4.2. *Proof of Proposition 1.* Using these lemmas, the proof of Proposition 1 is straightforward. Consider the choice of the regularization parameter, $\lambda_n = 16 \frac{2-\alpha}{\alpha} \sqrt{\frac{\log p}{n}}$. This choice satisfies the condition of Lemma 2, so that we may conclude that with probability greater than $1 - 2\exp(-c\lambda_n^2 n) \to 1$, we have

$$\|W^n\|_\infty \leq \frac{\alpha}{2-\alpha} \frac{\lambda}{4} \leq \frac{\lambda}{4}$$

using the fact that $\alpha \leq 1$. The remaining two conditions that we need to apply the technical lemmas concern upper bounds on the quantity $\lambda_n d$. In particular, for a sample size satisfying $n > \frac{100^2 D_{\max}^2}{C_{\min}^4} \frac{(2-\alpha)^4}{\alpha^4} d^2 \log p$, we have

$$\begin{aligned} \lambda_n d &= \frac{16(2-\alpha)}{\alpha} \sqrt{\frac{\log p}{n}} d \\ &\leq \frac{16 C_{\min}^2}{100 D_{\max}} \frac{\alpha}{(2-\alpha)} \\ &< \frac{C_{\min}^2}{10 D_{\max}} \end{aligned}$$

so that the conditions of both Lemmas 3 and 4 are satisfied.

We can now proceed to the proof of Proposition 1. Recalling our shorthand $Q^n = \nabla_\theta^2 \ell(\theta^*; \mathfrak{X}_1^n)$ and the fact that we have set $\widehat{\theta}_{S^c} = 0$ in our primal-dual construction, we can re-write condition (27) in block form as

$$Q_{S^c S}^n [\widehat{\theta}_S - \theta_S^*] = W_{S^c}^n - \lambda_n \widehat{z}_{S^c} + R_{S^c}^n, \tag{31a}$$

$$Q_{SS}^n [\widehat{\theta}_S - \theta_S^*] = W_S^n - \lambda_n \widehat{z}_S + R_S^n. \tag{31b}$$

Since the matrix Q_{SS}^n is invertible by assumption, the conditions (31) can be re-written as

$$Q_{S^c S}^n (Q_{SS}^n)^{-1} [W_S^n - \lambda_n \widehat{z}_S + R_S^n] = W_{S^c}^n - \lambda_n \widehat{z}_{S^c} + R_{S^c}^n. \tag{32}$$

Rearranging yields the condition,

$$[W_{S^c}^n - R_{S^c}^n] - Q_{S^c S}^n (Q_{SS}^n)^{-1} [W_S^n - R_S^n] + \lambda_n Q_{S^c S}^n (Q_{SS}^n)^{-1} \widehat{z}_S = \lambda_n \widehat{z}_{S^c}. \tag{33}$$

Strict dual feasibility. We now demonstrate that for the dual sub-vector $\widehat{z}_{S^c}$ defined by (33), we have $\|\widehat{z}_{S^c}\|_\infty < 1$. Using the triangle inequality and the mutual incoherence bound (14), we have that

$$\|\widehat{z}_{S^c}\|_\infty \leq |||Q^n_{S^cS}(Q^n_{SS})^{-1}|||_\infty \left[\frac{\|W^n_S\|_\infty}{\lambda_n} + \frac{\|R^n_S\|_\infty}{\lambda_n} + 1\right] + \frac{\|R^n_{S^c}\|_\infty}{\lambda_n} + \frac{\|W^n_{S^c}\|_\infty}{\lambda_n} \tag{34}$$

$$\leq (1-\alpha) + (2-\alpha)\left[\frac{\|R^n\|_\infty}{\lambda_n} + \frac{\|W^n\|_\infty}{\lambda_n}\right]. \tag{35}$$

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 (24) satisfies sign consistency, meaning that $\operatorname{sgn}(\widehat{\theta}_S) = \operatorname{sgn}(\theta^*_S)$. In order to do so, it suffices to show that

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

recalling the notation $\theta^*_{\min} := \min_{(r,t)\in E} |\theta^*_{rt}|$. From Lemma 3, we have $\|\theta_S - \theta^*_S\|_2 \leq \frac{5}{C_{\min}}\sqrt{d}\lambda_n$ so that

$$\frac{2}{\theta^*_{\min}}\|\theta_S - \theta^*_S\|_\infty \leq \frac{2}{\theta^*_{\min}}\|\theta_S - \theta^*_S\|_2 \leq \frac{2}{\theta^*_{\min}}\frac{5}{C_{\min}}\sqrt{d}\lambda_n,$$

which is less than one as long as $\theta^*_{\min} \geq \frac{10}{C_{\min}}\sqrt{d}\lambda_n$.

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., [11]) since the elements of Q^n are highly dependent. Recall the definitions

$$Q^* := \mathbb{E}_{\theta^*}[\eta(X;\theta^*)X_{\backslash r}X^T_{\backslash r}] \quad \text{and} \quad Q^n := \widehat{\mathbb{E}}[\eta(X;\theta^*)X_{\backslash r}X^T_{\backslash r}], \tag{36}$$

where $\mathbb{E}_{\theta^*}$ denotes the population expectation, and $\widehat{\mathbb{E}}$ denotes the empirical expectation, and the variance function η was defined previously in (12). The following lemma asserts that 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*

$$\mathbb{P}\left[\Lambda_{\max}\left[\frac{1}{n}\sum_{i=1}^{n} x_{\backslash r}^{(i)}(x_{\backslash r}^{(i)})^T\right] \geq D_{\max} + \delta\right] \leq 2\exp\left(-A\frac{\delta^2 n}{d^2} + B\log(d)\right), \tag{37a}$$

$$\mathbb{P}[\Lambda_{\min}(Q_{SS}^n) \leq C_{\min} - \delta] \leq 2\exp\left(-A\frac{\delta^2 n}{d^2} + B\log(d)\right). \tag{37b}$$

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 (14) 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*

$$\mathbb{P}\left[|||Q_{S^cS}^n(Q_{SS}^n)^{-1}|||_\infty \geq 1 - \frac{\alpha}{2}\right] \leq \exp\left(-K\frac{n}{d^3} + \log(p)\right). \tag{38}$$

Proofs of these two lemmas are provided in the following sections. Before proceeding, we take note of a simple bound to be used repeatedly throughout our arguments. By definition of the matrices $Q^n(\theta)$ and $Q(\theta)$ [see (20) and (11)], 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 $Z_{jk} = \frac{1}{n}\sum_{i=1}^n Z_{jk}^{(i)}$ where each $Z_{jk}^{(i)}$ is zero-mean and bounded (in particular, $|Z_{jk}^{(i)}| \leq 4$). By the Azuma–Hoeffding bound [15], for any indices $j, k = 1, \ldots, d$ and for any $\varepsilon > 0$, we have

$$\mathbb{P}[(Z_{jk})^2 \geq \varepsilon^2] = \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_{jk}^{(i)}\right| \geq \varepsilon\right] \leq 2\exp\left(-\frac{\varepsilon^2 n}{32}\right). \tag{39}$$

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 [16], we have

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

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

$$|||Q^n_{SS} - Q_{SS}|||_2 \leq \left(\sum_{j=1}^d \sum_{k=1}^d (Z_{jk})^2 \right)^{1/2}.$$

Setting $\varepsilon^2 = \delta^2/d^2$ in (39) and applying the union bound over the d^2 index pairs (j,k) then yields

$$\mathbb{P}[|||Q^n_{SS} - Q_{SS}|||_2 \geq \delta] \leq 2\exp\left(-K\frac{\delta^2 n}{d^2} + 2\log(d)\right). \tag{40}$$

Similarly, we have

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

which obeys the same upper bound (40) 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

$$T_1 := Q^*_{S^cS}[(Q^n_{SS})^{-1} - (Q^*_{SS})^{-1}], \tag{41a}$$

$$T_2 := [Q^n_{S^cS} - Q^*_{S^cS}](Q^*_{SS})^{-1}, \tag{41b}$$

$$T_3 := [Q^n_{S^cS} - Q^*_{S^cS}][(Q^n_{SS})^{-1} - (Q^*_{SS})^{-1}], \tag{41c}$$

$$T_4 := Q^*_{S^cS}(Q^*_{SS})^{-1}. \tag{41d}$$

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 (38), 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:*

$$\begin{aligned} &\mathbb{P}[\|\|Q^n_{S^cS} - Q^*_{S^cS}\|\|_\infty \geq \delta] \\ &\qquad \leq 2\exp\left(-K\frac{n\delta^2}{d^2} + \log(d) + \log(p-d)\right); \end{aligned} \tag{42a}$$

$$\begin{aligned} &\mathbb{P}[\|\|Q^n_{SS} - Q^*_{SS}\|\|_\infty \geq \delta] \\ &\qquad \leq 2\exp\left(-K\frac{n\delta^2}{d^2} + 2\log(d)\right); \end{aligned} \tag{42b}$$

$$\begin{aligned} &\mathbb{P}[\|\|(Q^n_{SS})^{-1} - (Q^*_{SS})^{-1}\|\|_\infty \geq \delta] \\ &\qquad \leq 4\exp\left(-K\frac{n\delta^2}{d^3} + K'\log(d)\right). \end{aligned} \tag{42c}$$

See Appendix C for the proof of these claims.

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

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

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

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

where we have used the incoherence assumption (A2). Using the bound (37b) 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 $1 - \exp(-Kn/d^2 + 2\log(d))$. Next, applying the bound (42b) with $\delta = c/\sqrt{d}$, we conclude that with probability greater than $1 - 2\exp(-Knc^2/d^3 + \log(d))$, we have

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

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

$$\mathbb{P}[\|\|T_1\|\|_\infty \geq \alpha/6] \leq 2\exp\left(-K\frac{nc^2}{d^3} + \log(d)\right). \tag{43}$$

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 (42a) with $\delta = \frac{\alpha}{3}\frac{C_{\min}}{\sqrt{d}}$ to conclude that

$$\mathbb{P}[|||T_2|||_\infty \geq \alpha/3] \leq 2\exp\left(-K\frac{n}{d^3} + \log(p-d)\right). \tag{44}$$

Control of third term. Finally, in order to bound the third term T_3, we apply the bounds (42a) and (42b), both with $\delta = \sqrt{\alpha/3}$, and use the fact that $\log(d) \leq \log(p-d)$ to conclude that

$$\mathbb{P}[|||T_3|||_\infty \geq \alpha/3] \leq 4\exp\left(-K\frac{n}{d^3} + \log(p-d)\right). \tag{45}$$

Putting together all of the pieces, we conclude that

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

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 by Koh, Kim and Boyd [20].

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 that $\theta^*_{st} = \pm\omega$ with equal probability or with *positive couplings* meaning that $\theta^*_{st} = \omega$ for all edges (s, t). In all cases, we set the regularization parameter λ_n as a constant factor of $\sqrt{\frac{\log p}{n}}$ as suggested by Theorem 1. 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

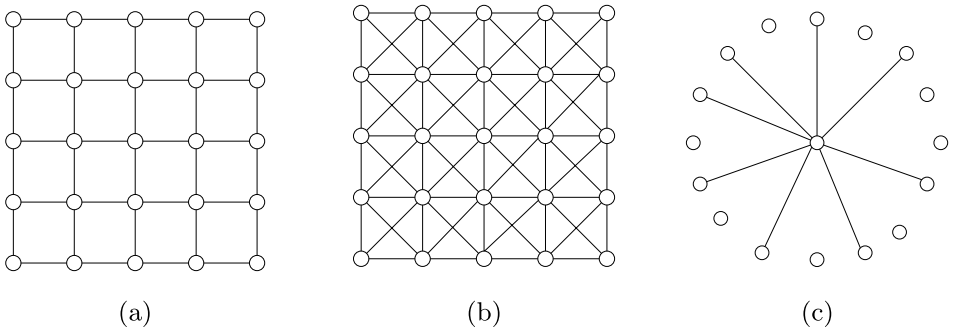

FIG. 1. *Illustrations of different graph classes used in simulations.* (a) *Four-nearest neighbor grid* ($d = 4$). (b) *Eight-nearest neighbor grid* ($d = 8$). (c) *Star-shaped graph* [$d = \Theta(p)$, *or* $d = \Theta(\log(p))$].

couplings [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 which is consistent with the prediction of Theorem 1. Figure 3 shows analogous results for the 8-nearest-neighbor lattice model ($d = 8$), for the same

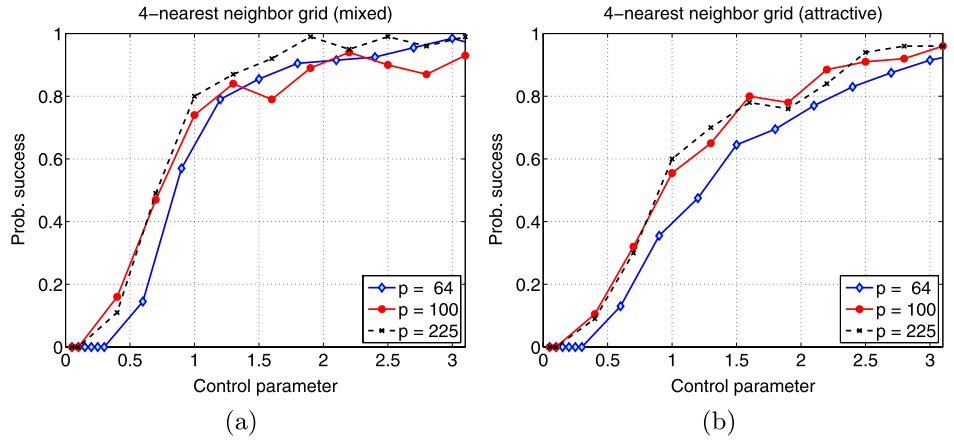

FIG. 2. *Plots of success probability* $\mathbb{P}[\widehat{\mathcal{N}}_{\pm}(r) = \mathcal{N}(r), \forall r]$ *versus the control parameter* $\beta(n, p, d) = n/[10d\log(p)]$ *for Ising models on 2-D grids with four nearest-neighbor interactions* ($d = 4$). (a) *Randomly chosen mixed sign couplings*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

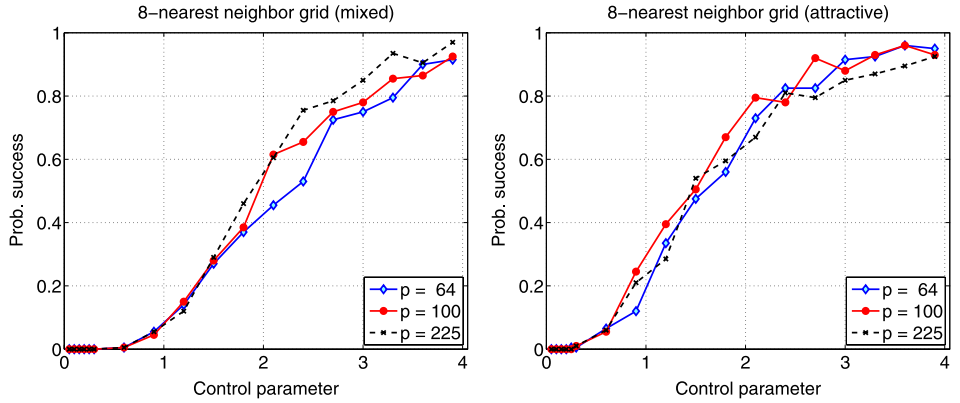

FIG. 3. *Plots of success probability* $\mathbb{P}[\widehat{\mathcal{N}}_{\pm}(r) = \mathcal{N}(r), \forall r]$ *versus the control parameter* $\beta(n,p,d) = n/[10d\log(p)]$ *for Ising models on 2-D grids with eight nearest-neighbor interactions* ($d = 8$). (a) *Randomly chosen mixed sign couplings*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

range of problem size $p \in \{64, 100, 225\}$ and for both mixed and attractive couplings. Notice how once again the curves for different problem sizes are all well aligned which is consistent with the prediction of Theorem 1.

For our next set of experiments, we investigate the performance of our method for a class of graphs with unbounded maximum degree d. In particular, we construct star-shaped graphs with p vertices by designating one

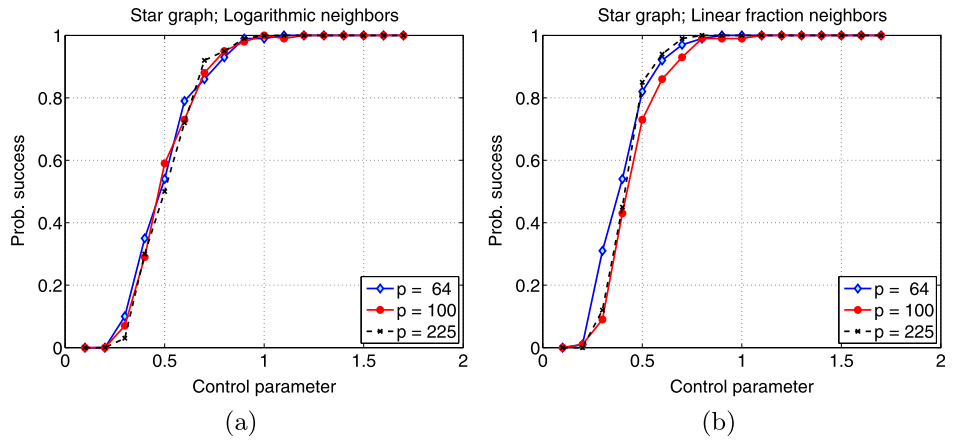

FIG. 4. *Plots of success probability* $\mathbb{P}[\widehat{\mathcal{N}}_{\pm}(r) = \mathcal{N}(r), \forall r]$ *versus the control parameter* $\beta(n,p,d) = n/[10d\log(p)]$ *for star-shaped graphs for attractive couplings with* (a) *logarithmic growth in degrees,* (b) *linear growth in degrees.*

node as the hub and connecting it to $d < (p-1)$ of its neighbors. For linear sparsity, we choose $d = \lceil 0.1p \rceil$, whereas for logarithmic sparsity we choose $d = \lceil \log(p) \rceil$. We again study 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.

Although the purpose of our experiments is mainly to illustrate the consequences of Theorem 1, we also include a comparison of our nodewise ℓ_1-penalized logistic regression-based method to two other graph estimation procedures. For the comparison, we use a star-shaped graph as in the previous plot, with one node designated as the hub connected to $d = \lceil 0.1p \rceil$ of its neighbors. It should be noted that among all graphs with a fixed total number of edges, this class of graphs is among the most difficult for our method to estimate. Indeed, the sufficient conditions of Theorem 1 scale logarithmically in the graph size p but polynomially in the maximum degree d; consequently, for a fixed total number of edges, our method requires the most samples when all the edges are connected to the same node, as in a star-shaped graph.

For comparative purposes, we also illustrate the performance of the PC algorithm of Spirtes, Glymour and Scheines [30] as well as the maximum weight tree method of Chow and Liu [7]. Since the star graph is a tree (cycle-free), both of these methods are applicable in this case. The PC algorithm is targeted to learning (equivalence classes of) directed acyclic graphs, and consists of two stages. In the first stage it starts from a completely connected undirected graph, and iteratively removes edges based on conditional independence tests so that at the end of this stage it is left with an undirected graph which is called a skeleton. In the second stage, it partially directs some of the edges in the skeleton so as to obtain a completed partially directed acyclic graph which corresponds to an equivalence class of directed acyclic graphs. As pointed out by Kalisch and Bühlmann [18], for high-dimensional problems, the output of the first stage, which is the undirected skeleton graph, could provide a useful characterization of the dependencies in the data. Following this suggestion, we use the skeleton graph determined by the first stage of the PC algorithm as an estimate of the graph structure. We use the `pcalg` R-package [18] as an implementation of the PC algorithm which uses partial correlations to test conditional independencies.

The Chow–Liu algorithm [7] is a method for exact maximum likelihood structure selection which is applicable to the case of trees. More specifically, it chooses, from among all trees with a specified number of edges, the tree that minimizes the Kullback–Leibler divergence to the empirical distribution defined by the samples. From an implementational point of view, it starts

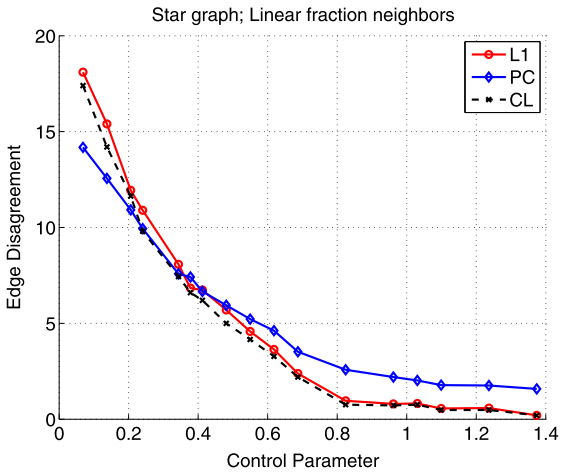

FIG. 5. *Plots of edge disagreement* $\mathbb{E}[|\{(s,t) \mid \widehat{E}_{st} \neq E^*_{st}\}|]$ *versus the control parameter* $\beta(n,p,d) = n/[10d\log(p)]$ *for star-shaped graphs where the hub node has degree* $d = \Theta(p)$. *The results here are shown for attractive couplings with* $\theta^*_{st} = 0.25$ *for all edges* (s,t) *belonging to the edge set. The* ℓ_1*-penalized logistic regression method (L1), the PC method (PC) and the maximum weight forest method of Chow and Liu (CL) are compared for* $p = 64$.

with a completely connected weighted graph with edge weights equal to the empirical mutual information between the incident node variables of the edge and then computes its maximum weight spanning tree. Since our underlying model is a star-shaped graph with fewer than $(p-1)$ edges, a spanning tree would necessarily include false positives. We thus estimate the maximum weight forest with d edges instead where we supplied the number of edges d in the true graph to the algorithm.

Figure 5 plots, for the three methods, the total number of edge disagreements between the estimated graphs and the true graph versus the control parameter $\beta = n/[10d\log(p)]$. Even though this class of graphs is especially challenging for a neighborhood-based method, the ℓ_1-penalized logistic-regression based method is competitive with the Chow–Liu algorithm, and except at very small sample sizes, it performs better than the PC algorithm for this problem.

7. Extensions to general discrete Markov random fields. Our method and analysis thus far has been specialized to the case of the binary pairwise Markov random fields. In this section, we briefly outline the extension to the case of general discrete pairwise Markov random fields. (Recall that for discrete Markov random fields, there is no loss of generality in assuming only pairwise interactions since by introducing auxiliary variables, higher-order interactions can be reformulated in a pairwise manner [34].) Let $X =$

$(X_1, \ldots, X_p)$ be a random vector, each variable X_i taking values in a set $\mathcal{X}$ of cardinality m, say $\mathcal{X} = \{1, 2, \ldots, m\}$. Let $G = (V, E)$ denote a graph with p nodes corresponding to the p variables $\{X_1, \ldots, X_p\}$, and let $\{\phi_s : \mathcal{X} \to \mathbb{R}, s \in V\}$ and $\{\phi_{st} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, (s,t) \in E\}$, respectively, denote collections of potential functions associated with the nodes and edges of the graph. These functions can be used to define a pairwise Markov random field over $(X_1, \ldots, X_p)$, with density

$$\mathbb{P}(x) \propto \exp\Biggl\{\sum_{s \in V} \phi_s(x_s) + \sum_{(s,t) \in E} \phi_{st}(x_s, x_t)\Biggr\}. \tag{46}$$

Since $\mathcal{X}$ is discrete, each potential function ϕ_{st} can be parameterized as linear combinations of $\{0,1\}$-valued indicator functions. In particular, for each $s \in V$ and $j \in \{1, \ldots, m-1\}$, we define

$$\mathbb{I}[x_s = j] = \begin{cases} 1, & \text{if } x_s = j, \\ 0, & \text{otherwise.} \end{cases}$$

Note we omit an indicator for $x_s = m$ from the list, since it is redundant given the indicators for $j = 1, \ldots, m-1$. In a similar fashion, we define the pairwise indicator functions $\mathbb{I}[x_s = j, x_t = k]$, for $(j,k) \in \{1, 2, \ldots, m-1\}^2$.

Any set of potential functions can then be written as

$$\phi_s(x_s) = \sum_{j \in \{1, \ldots, m-1\}} \theta^*_{s;j} \mathbb{I}[x_s = j] \qquad \text{for } s \in V,$$

and

$$\phi_{st}(x_s, x_t) = \sum_{(j,k) \in \{1, \ldots, m-1\}^2} \theta^*_{st;jk} \mathbb{I}[x_s = j, x_t = k] \qquad \text{for } (s,t) \in E.$$

Overall, the Markov random field can be parameterized in terms of the vector $\theta^*_s \in \mathbb{R}^{m-1}$ for each $s \in V$, and the vector $\theta^*_{st} \in \mathbb{R}^{(m-1)^2}$ associated with each edge. In discussing graphical model selection, it is convenient to associate a vector $\theta^*_{uv} \in \mathbb{R}^{(m-1)^2}$ to *every* pair of distinct vertices (u, v) with the understanding that $\theta^*_{uv} = 0$ if $(u, v) \notin E$.

With this set-up, we now describe a graph selection procedure that is the natural generalization of our procedure for the Ising model. As before we focus on recovering for each vertex $r \in V$ its neighborhood set and then combine the neighborhood sets across vertices to form the graph estimate.

For a binary Markov random field (1), there is a unique parameter θ^*_{rt} associated with each edge $(r,t) \in E$. For m-ary models, in contrast, there is a vector $\theta^*_{rt} \in \mathbb{R}^{(m-1)^2}$ of parameters associated with any edge (r,t). In order to describe a recovery procedure for the edges, let us define a matrix $\Theta^*_{\backslash r} \in \mathbb{R}^{(m-1)^2 \times (p-1)}$ where column u is given by the vector θ^*_{ru}. Note that unless vertex r is connected to all of its neighbors, many of the matrix

columns are zero. In particular, the problem of neighborhood estimation for vertex r corresponds to estimating the *column support* of the matrix $\Theta^*_{\backslash r}$—that is,

$$\mathcal{N}(r) = \{u \in V \setminus \{r\} \mid \|\theta^*_{ru}\|_2 \neq 0\}.$$

In order to estimate this column support, we consider the conditional distribution of X_r given the other variables $X_{\backslash\{r\}} = \{X_t | t \in V \setminus \{r\}\}$. For a binary model, this distribution is of the logistic form while for a general pairwise MRF, it takes the form

$$\mathbb{P}_\Theta[X_r = j \mid X_{\backslash r} = x_{\backslash r}] = \frac{\exp(\theta^*_{r;j} + \sum_{t\in V\setminus\{r\}} \sum_k \theta^*_{rt;jk} \mathbb{I}[x_t = k])}{\sum_\ell \exp(\theta^*_{r;\ell} + \sum_{t\in V\setminus\{r\}} \sum_k \theta^*_{rt;\ell k} \mathbb{I}[x_t = k])}. \tag{47}$$

Thus, X_r can be viewed as the response variable in a multiclass logistic regression in which the indicator functions associated with the other variables,

$$\{\mathbb{I}[x_t = k], t \in V \setminus \{r\}, k \in \{1, 2, \ldots, m-1\}\},$$

play the role of the covariates.

Accordingly, one method of recovering the row support of $\Theta^*_{\backslash r}$ is by performing multiclass logistic regression of X_r on the rest of the variables $X_{\backslash r}$ using a block ℓ_2/ℓ_1 penalty of the form

$$|||\Theta_{\backslash r}|||_{2,1} := \sum_{u\in V\setminus\{r\}} \|\theta_{ru}\|_2.$$

More specifically, let $\mathfrak{X}_1^n = \{x^{(1)}, \ldots, x^{(n)}\}$ denote an i.i.d. set of n samples, drawn from the discrete MRF (46). In order to estimate the neighborhood of node r, we solve the following convex program:

$$\min_{\Theta_{\backslash r}\in\mathbb{R}^{(m-1)^2\times(p-1)}} \{\ell(\Theta_{\backslash r}; \mathfrak{X}_1^n) + \lambda_n |||\Theta_{\backslash r}|||_{2,1}\}, \tag{48}$$

where $\ell(\Theta_{\backslash r}; \mathfrak{X}_1^n) := \frac{1}{n}\sum_{i=1}^n \log \mathbb{P}_\Theta[x_r^{(i)} \mid x_{\backslash r}^{(i)}]$ is the rescaled multiclass logistic likelihood defined by the conditional distribution (47), and $\lambda_n > 0$ is a regularization parameter.

The convex program (48) is the multiclass logistic analog of the group Lasso, a type of relaxation that has been studied in previous and on-going work on linear and logistic regression (e.g., [19, 22, 25, 38]). It should be possible to extend our analysis from the preceding sections so as to obtain similar high-dimensional consistency rates for this multiclass setting; the main difference is the slightly different sub-differential associated with the block ℓ_2/ℓ_1 norm. See Obozinski, Wainwright and Jordan [25] for some related work on support recovery using ℓ_2/ℓ_1 block-regularization for multivariate linear regression.

8. Conclusion. We have shown that a technique based on ℓ_1-regularized logistic regression can be used to perform consistent model selection in binary Ising 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. Simulation results show the accuracy of these theoretical predictions. 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 [29]. Overall, our experimental results are consistent with the conjecture that logistic regression procedure fails with high probability for sample sizes n that are smaller than $\mathcal{O}(d \log p)$. It would be interesting to prove such a converse result, to parallel the known upper and lower thresholds for success/failure of ℓ_1-regularized linear regression, or the Lasso (see [33]).

As discussed in Section 7, although the current analysis is applied to binary Markov random fields, the methods of this paper can be extended to general discrete graphical models with a higher number of states using a multinomial likelihood and some form of block regularization. It should also be possible and would be interesting to obtain high-dimensional rates in this setting. A final interesting direction for future work is the case of samples drawn in a non-i.i.d. manner from some unknown Markov random field; we suspect that similar results would hold for weakly dependent sampling schemes.

APPENDIX A: PROOF OF UNIQUENESS LEMMA

In this appendix, we prove Lemma 1. By Lagrangian duality, the penalized problem (7) 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.

By the definition of the ℓ_1-subdifferential, the subgradient vector $\widehat{z}$ can be expressed as a convex combination of sign vectors of the form

$$\widehat{z} = \sum_{v \in \{-1,+1\}^{p-1}} \alpha_v v, \tag{49}$$

where the weights α_v are nonnegative and sum to one. In fact, these weights correspond to an optimal vector of Lagrange multipliers for an alternative formulation of the problem in which α_v is the Lagrange multiplier for the constraint $\langle v, \theta \rangle \leq C(\lambda_n)$. From standard Lagrangian theory [3], it follows that any other optimal primal solution $\widetilde{\theta}$ must minimize the associated

Lagrangian—or equivalently, satisfy (21)—and moreover must satisfy the complementary slackness conditions $\alpha_v\{\langle v, \widetilde{\theta}\rangle - C\} = 0$ for all sign vectors v. But these conditions imply that $\langle \widehat{z}, \widetilde{\theta}\rangle = C = \|\widetilde{\theta}\|_1$ 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.

APPENDIX B: PROOFS FOR TECHNICAL LEMMAS

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

B.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

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

are zero-mean under $\mathbb{P}_{\theta^*}$, i.i.d. and bounded ($|Z_u^{(i)}| \leq 2$). Therefore, by the Azuma-Hoeffding inequality [15], we have, for any $\delta > 0$, $\mathbb{P}[|W_u^n| > \delta] \leq 2\exp(-\frac{n\delta^2}{8})$. Setting $\delta = \frac{\alpha\lambda_n}{4(2-\alpha)}$, we obtain

$$\mathbb{P}\left[\frac{2-\alpha}{\lambda_n}|W_u^n| > \frac{\alpha}{4}\right] \leq 2\exp\left(-\frac{\alpha^2\lambda_n^2}{128(2-\alpha)^2}n\right).$$

Finally, applying a union bound over the indices u of W^n yields

$$\mathbb{P}\left[\frac{2-\alpha}{\lambda_n}\|W^n\|_\infty > \frac{\alpha}{4}\right] \leq 2\exp\left(-\frac{\alpha^2\lambda_n^2}{128(2-\alpha)^2}n + \log(p)\right)$$

as claimed.

B.2. Proof of Lemma 3. Following a method used in a different context by Rothman et al. [28], we define the function $G : \mathbb{R}^d \to \mathbb{R}$ by

$$G(u_S) := \ell(\theta_S^* + u_S; \mathfrak{X}_1^n) - \ell(\theta_S^*; \mathfrak{X}_1^n) + \lambda_n(\|\theta_S^* + u_S\| - \|\theta_S^*\|). \tag{50}$$

It can be seen from (24) that $\widehat{u} = \widehat{\theta}_S - \theta_S^*$ minimizes G. Moreover, $G(0) = 0$ by construction; therefore, we must have $G(\widehat{u}) \leq 0$. Note also that G is convex. 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,

$$G(t\widehat{u} + (1-t)(0)) \le tG(\widehat{u}) + (1-t)G(0) \le 0,$$

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 $W = \nabla\ell(\theta^*; \mathfrak{X}_1^n)$, by a Taylor series expansion of the log likelihood component of G, we have

$$G(u) = W_S^T u + u^T[\nabla^2\ell(\theta_S^* + \alpha u)]u + \lambda_n(\|\theta_S^* + u_S\| - \|\theta_S^*\|) \tag{51}$$

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

$$|W_S^T u| \le \|W_S\|_\infty \|u\|_1 \le \|W_S\|_\infty \sqrt{d}\|u\|_2 \le (\lambda_n\sqrt{d})^2 \frac{M}{4}, \tag{52}$$

since $\|W_S\|_\infty \le \frac{\lambda_n}{4}$ by assumption.

Applying the triangle inequality to the last term in the expansion (51) yields

$$\lambda_n\|\theta_S^* + u_S\|_1 - \|\theta_S^*\|_1 \ge -\lambda_n\|u_S\|_1.$$

Since $\|u_S\|_1 \le \sqrt{d}\|u_S\|_2$, we have

$$\lambda_n\|\theta_S^* + u_S\|_1 - \|\theta_S^*\|_1 \ge -\lambda_n\sqrt{d}\|u_S\|_2 = -M(\sqrt{d}\lambda_n)^2. \tag{53}$$

Finally, turning to the middle Hessian term, we have

$$\begin{aligned} q^* &:= \Lambda_{\min}(\nabla^2\ell(\theta_S^* + \alpha u; \mathfrak{X}_1^n)) \\ &\ge \min_{\alpha\in[0,1]} \Lambda_{\min}(\nabla^2\ell(\theta_S^* + \alpha u_S; \mathfrak{X}_1^n)) \\ &= \min_{\alpha\in[0,1]} \Lambda_{\min}\left[\frac{1}{n}\sum_{i=1}^n \eta(x^{(i)}; \theta_S^* + \alpha u_S) x_S^{(i)} (x_S^{(i)})^T\right]. \end{aligned}$$

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

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

$$\geq C_{\min} - \max_{\alpha \in [0,1]} \Bigg\| \underbrace{\frac{1}{n} \sum_{i=1}^{n} \eta'(x^{(i)}; \theta_S^* + \alpha u_S)(\langle u_S, x_S^{(i)} \rangle) x_S^{(i)} (x_S^{(i)})^T}_{A(\alpha)} \Bigg\|_2 .$$

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

$$\begin{aligned} \langle y, A(\alpha) y \rangle &= \frac{1}{n} \sum_{i=1}^{n} \eta'(x^{(i)}; \theta_S^* + \alpha u_S)[\langle u_S, x_S^{(i)} \rangle][\langle x_S^{(i)}, y \rangle]^2 \\ &\leq \frac{1}{n} \sum_{i=1}^{n} |\eta'(x^{(i)}; \theta_S^* + \alpha u_S)| |\langle u_S, x_S^{(i)} \rangle| [\langle x_S^{(i)}, y \rangle]^2. \end{aligned}$$

Now note that $|\eta'(x^{(i)}; \theta_S^* + \alpha u_S)| \leq 1$, and

$$|\langle u_S, x_S^{(i)} \rangle| \leq \sqrt{d} \|u_S\|_2 = M \lambda_n d.$$

Moreover, we have

$$\frac{1}{n} \sum_{i=1}^{n} (\langle x_S^{(i)}, y \rangle)^2 \leq \Bigg\| \frac{1}{n} \sum_{i=1}^{n} x_S^{(i)} (x_S^{(i)})^T \Bigg\|_2 \leq D_{\max}$$

by assumption. Combining these pieces, we obtain

$$\max_{\alpha \in [0,1]} |||A(\alpha)|||_2 \leq D_{\max} M \lambda_n d \leq C_{\min}/2,$$

assuming that $\lambda_n \leq \frac{C_{\min}}{2MD_{\max}d}$. We verify this condition momentarily, after we have specified the constant M.

Under this condition, we have shown that

$$q^* := \Lambda_{\min}(\nabla^2 \ell(\theta_S^* + \alpha u; \mathfrak{X}_1^n)) \geq C_{\min}/2. \tag{54}$$

Finally, combining the bounds (52), (53), and (54) in the expression (51), we conclude that

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

This expression is strictly positive for $M = 5/C_{\min}$. Consequently, as long as

$$\lambda_n \leq \frac{C_{\min}}{2MD_{\max}d} = \frac{C_{\min}^2}{10 D_{\max} d}$$

as assumed in the statement of the lemma, we are guaranteed that

$$\|u_S\|_2 \leq M \lambda_n \sqrt{d} = \frac{5}{C_{\min}} \lambda_n \sqrt{d}$$

as claimed.

B.3. Proof of Lemma 4. We first show that the remainder term R^n satisfies the bound $\|R^n\|_\infty \le D_{\max}\|\widehat{\theta}_S - \theta^*_S\|_2^2$. Then the result of Lemma 3—namely, that $\|\widehat{\theta}_S - \theta^*_S\|_2 \le \frac{5}{C_{\min}}\lambda_n\sqrt{d}$—can be used to conclude that

$$\frac{\|R^n\|_\infty}{\lambda_n} \le \frac{25 D_{\max}}{C^2_{\min}}\lambda_n d$$

as claimed in Lemma 4.

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

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

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

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

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

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

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

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

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

APPENDIX C: PROOF OF LEMMA 7

Recall from the discussion leading up to the bound (39) 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

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

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

$$\begin{aligned}\mathbb{P}\left[\sum_{k\in S}|Z_{jk}| \geq \delta\right] &\leq \mathbb{P}[\exists k \in S||Z_{jk}| \geq \delta/d] \\ &\leq d\mathbb{P}[|Z_{jk}| \geq \delta/d];\end{aligned}$$

we then obtain

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

from which the claim (42a) follows by setting $\varepsilon = \delta/d$ in the Hoeffding bound (39). The proof of bound (42b) is analogous with the pre-factor $(p-d)$ replaced by d.

To prove the last claim (42c), we write

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

From the proof of Lemma 5, in particular equation (40), we have

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

for a constant B. Moreover, from (40), we have

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

so that the bound (42c) follows.

REFERENCES

[1] Abbeel, P., Koller, D. and Ng, A. Y. (2006). Learning factor graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.* **7** 1743–1788. MR2274423

[2] Banerjee, O., Ghaoui, L. E. and d'Asprémont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

[3] Bertsekas, D. (1995). *Nonlinear Programming*. Athena Scientific, Belmont, MA.

[4] Bresler, G., Mossel, E. and Sly, A. (2009). Reconstruction of Markov random fields from samples: Some easy observations and algorithms. Available at http://front.math.ucdavis.edu/0712.1402.

[5] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35** 2313–2351. MR2382644

[6] Chickering, D. (1995). Learning Bayesian networks is NP-complete. In *Learning from Data: Artificial Intelligence and Statistics V* (D. Fisher and H. Lenz, eds.). *Lecture Notes in Statistics* **112** 121–130. Springer, New York.

[7] Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* **14** 462–467.

[8] Cross, G. and Jain, A. (1983). Markov random field texture models. *IEEE Trans. PAMI* **5** 25–39.

[9] Csiszár, I. and Talata, Z. (2006). Consistent estimation of the basic neighborhood structure of Markov random fields. *Ann. Statist.* **34** 123–145. MR2275237

[10] Dasgupta, S. (1999). Learning polytrees. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI-99)*. Morgan Kaufmann, San Francisco, CA.

[11] Davidson, K. R. and Szarek, S. J. (2001). Local operator theory, random matrices, and Banach spaces. In *Handbook of the Geometry of Banach Spaces* **1** 317–336. Elsevier, Amsterdam. MR1863696

[12] Donoho, D. and Elad, M. (2003). Maximal sparsity representation via ℓ_1 minimization. *Proc. Natl. Acad. Sci. USA* **100** 2197–2202. MR1963681

[13] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI* **6** 721–741.

[14] Hassner, M. and Sklansky, J. (1980). The use of Markov random fields as models of texture. *Comp. Graphics Image Proc.* **12** 357–370.

[15] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[16] Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR0832183

[17] Ising, E. (1925). Beitrag zur theorie der ferromagnetismus. *Zeitschrift für Physik* **31** 253–258.

[18] Kalisch, M. and Buhlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

[19] Kim, Y., Kim, J. and Kim, Y. (2005). Blockwise sparse regression. *Statist. Sinica* **16** 375–390.

[20] Koh, K., Kim, S. J. and Boyd, S. (2007). An interior-point method for large-scale ℓ_1-regularized logistic regression. *J. Mach. Learn. Res.* **3** 1519–1555. MR2332440

[21] Manning, C. D. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA. MR1722790

[22] Meier, L., van de Geer, S. and Bühlmann, P. (2007). The group lasso for logistic regression. Technical report, Mathematics Dept., Swiss Federal Institute of Technology Zürich.

[23] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[24] Ng, A. Y. (2004). Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML-04)*. Morgan Kaufmann, San Francisco, CA.

[25] Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2008). Union support recovery in high-dimensional multivariate regression. Technical report, Dept. Statistics, Univ. California, Berkeley.

[26] Ripley, B. D. (1981). *Spatial Statistics*. Wiley, New York. MR0624436

[27] Rockafellar, G. (1970). *Convex Analysis*. Princeton Univ. Press, Princeton. MR0274683

[28] Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

[29] Santhanam, N. P. and Wainwright, M. J. (2008). Information-theoretic limits of high-dimensional graphical model selection. In *International Symposium on Information Theory*. Toronto, Canada.

[30] Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction and Search*. MIT Press, Cambridge, MA. MR1815675

[31] Srebro, N. (2003). Maximum likelihood bounded tree-width Markov networks. *Artificial Intelligence* **143** 123–138. MR1963987

[32] Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals. *IEEE Trans. Inform. Theory* **51** 1030–1051. MR2238069

[33] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202.

[34] Wainwright, M. J. and Jordan, M. I. (2003). Graphical models, exponential families, and variational inference. Technical Report 649, Dept. Statistics, Univ. California, Berkeley. MR2082153

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

[36] Welsh, D. J. A. (1993). *Complexity: Knots, Colourings, and Counting*. Cambridge Univ. Press, Cambridge. MR1245272

[37] Woods, J. (1978). Markov image modeling. *IEEE Trans. Automat. Control* **23** 846–850.

[38] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

[39] Zhao, P. and Yu, B. (2007). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7** 2541–2567. MR2274449

P. Ravikumar
M. J. Wainwright
Department of Statistics
University of California
Berkeley, California 94720
USA
E-mail: pradeepr@stat.berkeley.edu
wainwrig@stat.berkeley.edu

J. D. Lafferty
Computer Science Department
and Machine Learning Department
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
USA
E-mail: lafferty@cs.cmu.edu