Presented at NIPS-2006 conference
December 2006; Vancouver, BC, Canada

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

Martin J. Wainwright
Department of Statistics
Department of EECS
Univ. of California, Berkeley
Berkeley, CA 94720

Pradeep Ravikumar
Machine Learning Dept.
Carnegie Mellon Univ.
Pittsburgh, PA 15213

John D. Lafferty
Computer Science Dept.
Machine Learning Dept.
Carnegie Mellon Univ.
Pittsburgh, PA 15213

Abstract

We focus on 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 result is to establish 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 mutual incoherence conditions analogous to those imposed in previous work on linear regression, we prove that consistent neighborhood selection can be obtained as long as the number of observations n grows more quickly than $6d^6 \log d + 2d^5 \log p$, thereby establishing that logarithmic growth in the number of samples n relative to graph size p is sufficient to achieve neighborhood consistency.

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

1 Introduction

Consider a p-dimensional discrete random variable $X = (X_1, X_2, \ldots, X_p)$ where the distribution of X is governed by an unknown undirected graphical model. In this paper, we investigate the problem of estimating the graph structure from an i.i.d. sample of n data points $\{x^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)}\}_{i=1}^n$. This structure learning problem plays an important role in a broad range of applications where graphical models are used as a probabilistic representation tool, including image processing, document analysis and medical diagnosis. Our approach is to perform an ℓ_1-regularized logistic regression of each variable on the remaining variables, and to use the sparsity pattern of the regression vector to infer the underlying neighborhood structure. The main contribution of the paper is a theoretical analysis showing that, under suitable conditions, this procedure recovers the true graph structure with probability one, in the high-dimensional setting in which both the sample size n and graph size $p = p(n)$ increase to infinity.

The problem of structure learning for discrete graphical models—due to both its importance and difficulty—has attracted considerable attention. Constraint based approaches use hypothesis testing to estimate the set of conditional independencies in the data, and then determine a graph that most closely represents those independencies [8]. An alternative approach is to view the problem as estimation of a stochastic model, combining a scoring metric

on candidate graph structures with a goodness of fit measure to the data. The scoring metric approach must be used together with a search procedure that generates candidate graph structures to be scored. The combinatorial space of graph structures is super-exponential, however, and Chickering [1] shows that this problem is in general NP-hard. The space of candidate structures in scoring based approaches is typically restricted to directed models (Bayesian networks) since the computation of typical score metrics involves computing the normalization constant of the graphical model distribution, which is intractable for general undirected models. Estimation of graph structures in undirected models has thus largely been restricted to simple graph classes such as trees [2], polytrees [3] and hypertrees [9].

The technique of ℓ_1 regularization for estimation of sparse models or signals has a long history in many fields; we refer to Tropp [10] for a recent survey. A surge of recent work has shown that ℓ_1-regularization can lead to practical algorithms with strong theoretical guarantees (e.g., [4, 5, 6, 10, 11, 12]). In this paper, we adapt the technique of ℓ_1-regularized logistic regression to the problem of inferring graph structure. The technique is computationally efficient and thus well-suited to high dimensional problems, since it involves the solution only of standard convex programs. Our main result establishes conditions on the sample size n, graph size p and maximum neighborhood size d under which the true neighborhood structure can be inferred with probability one as (n, p, d) increase. Our analysis, though asymptotic in nature, leads to growth conditions that are sufficiently weak so as to require only that the number of observations n grow logarithmically in terms of the graph size. Consequently, our results establish that graphical structure can be learned from relatively sparse data. Our analysis and results are similar in spirit to the recent work of Meinshausen and Bühlmann [5] on covariance selection in Gaussian graphical models, but focusing rather on the case of discrete models.

The remainder of this paper is organized as follows. In Section 2, we formulate the problem and establish notation, before moving on to a precise statement of our main result, and a high-level proof outline in Section 3. Sections 4 and 5 detail the proof, with some technical details deferred to the full-length version. Finally, we provide experimental results and a concluding discussion in Section 6.

2 Problem Formulation and Notation

Let $G = (V, E)$ denote a graph with vertex set V of size $|V| = p$ and edge set E. We denote by $\mathcal{N}(s)$ the set of neighbors of a vertex $v \in V$; that is $\mathcal{N}(s) = \{(s,t) \in E\}$. A pairwise graphical model with graph G is a family of probability distributions for a random variable $X = (X_1, X_2, \ldots, X_p)$ given by $p(x) \propto \prod_{(s,t)\in E} \psi_{st}(x_s, x_t)$. In this paper, we restrict our attention to the case where each $x_s \in \{0, 1\}$ is binary, and the family of probability distributions is given by the Ising model

$$p(x; \theta) = \exp\left(\sum_{s\in V} \theta_s x_s + \sum_{(s,t)\in E} \theta_{st} x_s x_t - \Psi(\theta)\right). \tag{1}$$

Given such an exponential family in a minimal representation, the log partition function $\Psi(\theta)$ is strictly convex, which ensures that the parameter matrix θ is identifiable.

We address the following problem of graph learning. Given n samples $x^{(i)} \in \{0, 1\}^p$ drawn from an unknown distribution $p(x; \theta^*)$ of the form (1), let $\widehat{E}_n$ be an estimated set of edges. Our set-up includes the important situation in which the number of variables p may be large relative to the sample size n. In particular, we allow the graph $G_n = (V_n, E_n)$ to vary with n, so that the number of variables $p = |V_n|$ and the sizes of the neighborhoods $d_s := |\mathcal{N}(s)|$ may vary with sample size. (For notational clarity we will sometimes omit subscripts indicating a dependence on n.) The goal is to construct an estimator $\widehat{E}_n$ for which $\mathbb{P}[\widehat{E}_n = E_n] \to 1$ as $n \to \infty$. Equivalently, we consider the problem of estimating neighborhoods $\widehat{\mathcal{N}}_n(s) \subset V_n$ so that $\mathbb{P}[\widehat{\mathcal{N}}_n(s) = \mathcal{N}(s),\ \forall\, s \in V_n] \longrightarrow 1$. For many problems of interest, the graphical model provides a compact representation where the size of the neighborhoods are typically small—say $d_s \ll p$ for all $s \in V_n$. Our goal is to use ℓ_1-regularized logistic regression to estimate these neighborhoods; for this paper, the actual values of the parameters θ_{ij} is a secondary concern.

Given input data $\{(z^{(i)}, y^{(i)})\}$, where $z^{(i)}$ is a p-dimensional covariate and $y^{(i)} \in \{0,1\}$ is a binary response, logistic regression involves minimizing the negative log likelihood

$$f_s(\theta; x) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(1 + \exp(\theta^T z^{(i)})) - y^{(i)} \theta^T z^{(i)} \right\}. \tag{2}$$

We focus on regularized version of this regression problem, involving an ℓ_1 constraint on (a subset of) the parameter vector θ. For convenience, we assume that $z_1^{(i)} = 1$ is a constant so that θ_1 is a bias term, which is not regularized; we denote by $\theta_{\backslash s}$ the vector of all coefficients of θ except the one in position s. For the graph learning task, we regress each variable X_s onto the remaining variables, sharing the same data $x^{(i)}$ across problems. This leads to the following collection of optimization problems (p in total, one for each graph node):

$$\widehat{\theta}^{s,\lambda} = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\log(1 + \exp(\theta^T z^{(i,s)})) - x_s^{(i)} \theta^T z^{(i,s)} \right] + \lambda_n \|\theta_{\backslash s}\|_1 \right\}. \tag{3}$$

where $s \in V$, and $z^{(i,s)} \in \{0,1\}^p$ denotes the vector where $z_t^{(i,s)} = x_t^{(i)}$ for $t \neq s$ and $z_s^{(i,s)} = 1$. The parameter θ_s acts as a bias term, and is not regularized. Thus, the quantity $\widehat{\theta}_t^{s,\lambda}$ can be thought of as a penalized conditional likelihood estimate of $\theta_{s,t}$. Our estimate of the neighborhood $\mathcal{N}(s)$ is then given by

$$\widehat{\mathcal{N}}_n(s) = \left\{ t \in V,\ t \neq s\ :\ \widehat{\theta}_t^{s,\lambda} \neq 0 \right\}.$$

Our goal is to provide conditions on the graphical model—in particular, relations among the number of nodes p, number of observations n and maximum node degree d—that ensure that the collection of neighborhood estimates (2), one for each node s of the graph, is consistent with high probability.

We conclude this section with some additional notation that is used throughout the sequel. Defining the probability $p(z^{(i,s)}; \theta) := [1 + \exp(-\theta^T z^{(i,s)})]^{-1}$, straightforward calculations yield the gradient and Hessian, respectively, of the negative log likelihood (2):

$$\nabla_\theta f_s(\theta; x) = \frac{1}{n} \sum_{i=1}^{n} p(z^{(i,s)}; \theta)\, z^{(i,s)} - \theta^T \left(\frac{1}{n} \sum_{i=1}^{n} x_s^{(i)} z^{(i,s)} \right) \tag{4a}$$

$$\nabla_\theta^2 f_s(\theta; x) = \frac{1}{n} \sum_{i=1}^{n} p(z^{(i,s)}; \theta)\, [1 - p(z^{(i,s)}; \theta)]\, z^{(i,s)}\, (z^{(i,s)})^T. \tag{4b}$$

Finally, for ease of notation, we make frequent use the shortand $Q_s(\theta) = \nabla^2 f_s(\theta; x)$.

3 Main Result and Outline of Analysis

In this section, we begin with a precise statement of our main result, and then provide a high-level overview of the key steps involved in its proof.

3.1 Statement of main result

We begin by stating the assumptions that underlie our main result. A subset of the assumptions involve the Fisher information matrix associated with the logistic regression model, defined for each node $s \in V$ as

$$Q_s^* = \mathbb{E}\left[p_s(Z; \theta^*)\, \{1 - p_s(Z; \theta^*)\} Z Z^T \right], \tag{5}$$

Note that Q_s^* is the population average of the Hessian $Q_s(\theta^*)$. For ease of notation we use S to denote the neighborhood $\mathcal{N}(s)$, and S^c to denote the complement $V - \mathcal{N}(s)$. Our first two assumptions (A1 and A2) place restrictions on the dependency and coherence structure of this Fisher information matrix. We note that these first two assumptions are analogous to conditions imposed in previous work [5, 10, 11, 12] on linear regression. Our third assumption is a growth rate condition on the triple (n, p, d).

[A1] Dependency condition: We require that the subset of the Fisher information matrix corresponding to the relevant covariates has bounded eigenvalues: namely, there exist constants $C_{min} > 0$ and $C_{max} < +\infty$ such that

$$C_{min} \ \leq \ \Lambda_{min}(Q^*_{SS}), \qquad and \qquad \Lambda_{max}(Q^*_{SS}) \ \leq \ C_{max}. \tag{6}$$

These conditions ensure that the relevant covariates do not become overly dependent, and can be guaranteed (for instance) by assuming that $\widehat{\theta}^{s,\lambda}$ lies within a compact set.

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

$$\|Q^*_{S^cS}(Q^*_{SS})^{-1}\|_\infty \quad \leq \quad 1 - \epsilon. \tag{7}$$

Analogous conditions are required for the success of the Lasso in the case of linear regression [5, 10, 11, 12].

[A3] Growth rates: Our second set of assumptions involve the growth rates of the number of observations n, the graph size p, and the maximum node degree d. In particular, we require that:

$$\frac{n}{d^5} - 6d\log(d) - 2\log(p) \quad \rightarrow \quad +\infty. \tag{8}$$

Note that this condition allows the graph size p to grow exponentially with the number of observations (i.e., $p(n) = \exp(n^\alpha)$ for some $\alpha \in (0, 1)$. Moreover, it is worthwhile noting that for model selection in graphical models, one is typically interested in node degrees d that remain bounded (e.g., $d = O(1)$), or grow only weakly with graph size (say $d = o(\log p)$).

With these assumptions, we now state our main result:

Theorem 1. *Given a graphical model and triple (n, p, d) such that conditions A1 through A3 are satisfied, suppose that the regularization parameter λ_n is chosen such that (a) $n\lambda_n^2 - 2\log(p) \rightarrow +\infty$, and (b) $d\lambda_n \rightarrow 0$. Then $\mathbb{P}[\widehat{\mathcal{N}}_n(s) = \mathcal{N}(s), \ \forall\, s \in V_n] \rightarrow 1$ as $n \rightarrow +\infty$.*

3.2 Outline of analysis

We now provide a high-level roadmap of the main steps involved in our proof of Theorem 1. Our approach is based on the notion of a *primal witness*: in particular, focusing our attention on a fixed node $s \in V$, we define a constructive procedure for generating a primal vector $\widehat{\theta} \in \mathbb{R}^p$ as well as a corresponding subgradient $\widehat{z} \in \mathbb{R}^n$ that together satisfy the zero-subgradient optimality conditions associated with the convex program (3). We then show that this construction succeeds with probability converging to one under the stated conditions. A key fact is that the convergence rate is sufficiently fast that a simple union bound over all graph nodes shows that we achieve consistent neighborhood estimation for all nodes simultaneously.

To provide some insight into the nature of our construction, the analysis in Section 4 shows the neighborhood $\mathcal{N}(s)$ is correctly recovered if and only if the pair $(\widehat{\theta}, \widehat{z})$ satisfies the following four conditions: (a) $\widehat{\theta}_{S^c} = 0$; (b) $|\widehat{\theta}_t| > 0 \quad$ for all $t \in S$; (c) $\widehat{z}_S = \mathrm{sgn}(\theta^*_S)$; and (d) $\|\widehat{z}_{S^c}\|_\infty < 1$. The first step in our construction is to choose the pair $(\widehat{\theta}, \widehat{z})$ such that both conditions (a) and (c) hold. The remainder of the analysis is then devoted to establishing that properties (b) and (d) hold with high probability.

In the first part of our analysis, we assume that the dependence (A1) mutual incoherence (A2) conditions hold for the *sample Fisher information matrices* $Q_s(\theta^*)$ defined below equation (4b). Under this assumption, we then show that the conditions on λ_n in the theorem statement suffice to guarantee that properties (b) and (d) hold for the constructed pair $(\widehat{\theta}, \widehat{z})$. The remainder of the analysis, provided in the full-length version of this paper, is

devoted to showing that under the specified growth conditions (A3), imposing incoherence and dependence assumptions on the *population version* of the Fisher information $Q^*(\theta^*)$ guarantees (with high probability) that analogous conditions hold for the sample quantities $Q_s(\theta^*)$. While it follows immediately from the law of large numbers that the empirical Fsiher information $Q^n_{AA}(\theta^*)$ converges to the population version Q^*_{AA} for any *fixed* subset, the delicacy is that we require controlling this convergence over subsets of increasing size. Our analysis therefore requires the use of uniform laws of large numbers [7].

4 Primal-Dual Relations for ℓ_1-Regularized Logistic Regression

Basic convexity theory can be used to characterize the solutions of ℓ_1-regularized logistic regression. We assume in this section that θ_1 corresponds to the unregularized bias term, and omit the dependence on sample size n in the notation. The objective is to compute

$$\min_{\theta\in\mathbb{R}^p} \mathcal{L}(\theta,\lambda) = \min_{\theta\in\mathbb{R}^p} \left\{ f(\theta;x) + \lambda\left(\|\theta_{\setminus 1}\|_1 - b\right)\right\} = \min_{\theta\in\mathbb{R}^p} \left\{ f(\theta;x) + \lambda\|\theta_{\setminus 1}\|_1\right\} \quad (9)$$

The function $\mathcal{L}(\theta,\lambda)$ is the Lagrangian function for the problem of minimizing $f(\theta;x)$ subject to $\|\theta_{\setminus 1}\|_1 \le b$ for some b. The dual function is $h(\lambda) = \inf_\theta \mathcal{L}(\theta,\lambda)$.

If $p \le n$ then $f(\theta;x)$ is a strictly convex function of θ. Since the ℓ_1-norm is convex, it follows that $\mathcal{L}(\theta,\lambda)$ is convex in θ and strictly convex in θ for $p \le n$. Therefore the set of solutions to (9) is convex. If $\widehat{\theta}$ and $\widehat{\theta}'$ are two solutions, then by convexity $\widehat{\theta} + \rho(\widehat{\theta}' - \widehat{\theta})$ is also a solution for any $\rho \in [0,1]$. Since the solutions minimize $f(\theta;x)$ subject to $\|\theta_{\setminus 1}\|_1 \le b$, the value of $f(\widehat{\theta} + \rho(\widehat{\theta}' - \widehat{\theta}))$ is independent of ρ, and $\nabla_\theta f(\widehat{\theta};x)$ is independent of the particular solution $\widehat{\theta}$. These facts are summarized below.

Lemma 1. *If $p \le n$ then a unique solution to (9) exists. If $p \ge n$ then the set of solutions is convex, with the value of $\nabla_\theta f(\widehat{\theta};x)$ constant across all solutions. In particular, if $p \ge n$ and $|\nabla_{\theta_t} f(\widehat{\theta};x)| < \lambda$ for some solution $\widehat{\theta}$, then $\widehat{\theta}_t = 0$ for all solutions.*

The subgradient $\partial\|\theta_{\setminus 1}\|_1 \subset \mathbb{R}^p$ is the collection of all vectors z satisfying $|z_t| \le 1$ and

$$z_t = \begin{cases} 0 & \text{for } t = 1 \\ \text{sign}(\theta_t) & \text{if } \theta_t \neq 0. \end{cases}$$

Any optimum of (9) must satisfy

$$\partial_\theta \mathcal{L}(\widehat{\theta},\lambda) = \nabla_\theta f(\widehat{\theta};x) + \lambda z = 0 \quad (10)$$

for some $z \in \partial\|\theta_{\setminus 1}\|$. The analysis in the following sections shows that, with high probability, a primal-dual pair $(\widehat{\theta},\widehat{z})$ can be constructed so that $|\widehat{z}_t| < 1$ and therefore $\widehat{\theta}_t = 0$ in case $\theta^*_t = 0$ in the true model θ^* from which the data are generated.

5 Constructing a Primal-Dual Pair

We now fix a variable X_s for the logistic regression, denoting the set of variables in its neighborhood by S. From the results of the previous section we observe that the ℓ_1-regularized regression recovers the sparsity pattern if and only if there exists a primal-dual solution pair $(\widehat{\theta},\widehat{z})$ satisfying the zero-subgradient condition, and the conditions (a) $\widehat{\theta}_{S^c} = 0$; (b) $|\widehat{\theta}_t| > 0 \quad$ for all $t \in S$ and $\text{sgn}(\widehat{\theta}_S) = \text{sgn}(\theta^*{}_S)$; (c) $\widehat{z}_S = \text{sgn}(\theta^*_S)$; and (d) $\|\widehat{z}_{S^c}\|_\infty < 1$.

Our proof proceeds by showing the existence (with high probability) of a primal-dual pair $(\widehat{\theta},\widehat{z})$ that satisfy these conditions. We begin by setting $\widehat{\theta}_{S^c} = 0$, so that (a) holds, and also setting $\widehat{z}_S = \text{sgn}(\widehat{\theta}_S)$, so that (c) holds. We first establish a consistency result when incoherence conditions are imposed on the sample Fisher information Q^n. The remaining analysis, deferred to the full-length version, establishes that the incoherence assumption (A2) on the population version ensures that the sample version also obeys the property with probability converging to one exponentially fast.

Theorem 2. *Suppose that*

$$\|Q^n_{S^cS}(Q^n_{SS})^{-1}\|_\infty \quad \leq \quad 1-\epsilon \tag{11}$$

for some $\epsilon \in (0,1]$. *Assume that* $\lambda_n \to 0$ *is chosen that* $\lambda_n^2 n - \log(p) \to +\infty$ *and* $\lambda_n d \to 0$. *Then* $\mathbb{P}\left(\widehat{\mathcal{N}}(s) = \mathcal{N}(s)\right) = 1 - O(\exp(-cn^\gamma))$ *for some* $\gamma > 0$.

Proof. Let us introduce the notation

$$W^n \quad := \quad \frac{1}{n}\sum_{i=1}^n z^{(i,s)}\left(x_s^{(i)} - \frac{\exp(\theta^{*T} z^{(i,s)})}{1+\exp(\theta^{*T} z^{(i,s)})}\right)$$

Substituting into the subgradient optimality condition (10) yields the equivalent condition

$$\nabla f(\widehat{\theta}; x) - \nabla f(\theta; x) - W^n + \lambda_n \widehat{z} \quad = \quad 0. \tag{12}$$

By a Taylor series expansion, this condition can be re-written as

$$\nabla^2 f(\theta^*; x)\,[\widehat{\theta} - \theta^*] \quad = \quad W^n - \lambda_n \widehat{z} + R^n, \tag{13}$$

where the remainder R^n is a term of order $\|R^n\|_2 = O(\|\widehat{\theta} - \theta^*\|^2)$.

Using our shorthand $Q^n = \nabla^2_\theta f(\theta^*; x)$, we write the zero-subgradient condition (13) in block form as:

$$Q^n_{S^cS}\,[\widehat{\theta}^{s,\lambda}_S - \theta^*_S] \quad = \quad W^n_{S^c} - \lambda_n \widehat{z}_{S^c} + R^n_{S^c}, \tag{14a}$$

$$Q^n_{SS}\,[\widehat{\theta}^{s,\lambda}_S - \theta^*_S] \quad = \quad W^n_S - \lambda_n \widehat{z}_S + R^n_S. \tag{14b}$$

It can be shown that the matrix Q^n_{SS} is invertible w.p. one, so that these conditions can be rewritten as

$$Q^n_{S^cS}\,(Q^n_{SS})^{-1}\,[W^n_S - \lambda_n \widehat{z}_S + R^n_S] \quad = \quad W^n_{S^c} - \lambda_n \widehat{z}_{S^c} + R^n_{S^c}. \tag{15}$$

Re-arranging yields the condition

$$Q^n_{S^cS}\,(Q^n_{SS})^{-1}\,[W^n_S - R^n_S] - [W^n_{S^c} - R^n_{S^c}] + \lambda_n Q^n_{S^cS}\,(Q^n_{SS})^{-1}\widehat{z}_S \quad = \quad \lambda_n \widehat{z}_{S^c}. \tag{16}$$

Analysis of condition (d): We now demonstrate that $\|\widehat{z}_{S^c}\|_\infty < 1$. Using triangle inequality and the sample incoherence bound (11) we have that

$$\|\widehat{z}_{S^c}\|_\infty \quad \leq \quad \frac{(2-\epsilon)}{\lambda_n}\,[\|W^n\|_\infty + \|R^n\|_\infty] + (1-\epsilon) \tag{17}$$

We complete the proof that $\|\widehat{z}_{S^c}\|_\infty < 1$ with the following two lemmas, proved in the full-length version.

Lemma 2. *If* $n\lambda_n^2 - \log(p) \to +\infty$, *then*

$$\mathbb{P}\left(\frac{2-\epsilon}{\lambda_n}\|W^n\|_\infty \geq \frac{\epsilon}{4}\right) \quad \to \quad 0 \tag{18}$$

at rate $O(\exp\left(-n\lambda_n^2 + \log(p)\right))$.

Lemma 3. *If* $n\lambda_n^2 - \log(p) \to +\infty$ *and* $d\lambda_n \to 0$, *then we have*

$$\mathbb{P}\left(\frac{2-\epsilon}{\lambda_n}\|R^n\|_\infty \geq \frac{\epsilon}{4}\right) \quad \to \quad 0 \tag{19}$$

at rate $O(\exp\left(-n\lambda_n^2 + \log(p)\right))$.

We apply these two lemmas to the bound (17) to obtain that with probability converging to one at rate $O(\exp\left\{\exp\left(n\lambda_n^2 - \log(p)\right)\right\})$, we have

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

Analysis of condition (b): We next show that condition (b) can be satisfied, so that $\mathrm{sgn}(\widehat{\theta}_S) = \mathrm{sgn}(\theta^*{}_S)$. Define $\rho_n := \min_{i \in S} |\theta^*_S|$. From equation (14b), we have

$$\widehat{\theta}^{s,\lambda}_S \;\; = \;\; \theta^*_S - (Q^n_{SS})^{-1} \; [W_S - \lambda_n \widehat{z}_S + R_S] \, . \tag{20}$$

Therefore, in order to establish that $|\widehat{\theta}^{s,\lambda}_i| > 0$ for all $i \in S$, and moreover that $\mathrm{sign}(\widehat{\theta}^{s,\lambda}_S) = \mathrm{sign}(\theta^*_S)$, it suffices to show that

$$\left\| (Q^n_{SS})^{-1} \; [W_S - \lambda_n \widehat{z}_S + R_S] \right\|_\infty \;\; \leq \;\; \frac{\rho_n}{2}.$$

Using our eigenvalue bounds, we have

$$\begin{aligned} \left\| (Q^n_{SS})^{-1} \; [W_S - \lambda_n \widehat{z}_S + R_S] \right\|_\infty & \leq \|(Q^n_{SS})^{-1}\|_\infty \; [\|W_S\|_\infty + \lambda_n + \|R_S\|_\infty] \\ & \leq \sqrt{d} \, \|(Q^n_{SS})^{-1}\|_2 \, [\|W_S\|_\infty + \lambda_n + \|R_S\|_\infty] \\ & \leq \frac{\sqrt{d}}{C_{min}} \, [\|W_S\|_\infty + \lambda_n + \|R_S\|_\infty] \, . \end{aligned}$$

In fact, the righthand side tends to zero from our earlier results on W and R, and the assumption that $\lambda_n d \to 0$. Together with the exponential rates of convergence established by the stated lemmas, this completes the proof of the result.

6 Experimental Results

We briefly describe some experimental results that demonstrate the practical viability and performance of our proposed method. We generated random Ising models (1) using the following procedure: for a given graph size p and maximum degree d, we started with a graph with disconnected cliques of size less than or equal to ten, and for each node, removed edges randomly until the sparsity condition (degree less than d) was satisfied. For all edges (s,t) present in the resulting random graph, we chose the edge weight $\theta_{st} \sim \mathcal{U}[-3,3]$. We drew n i.i.d. samples from the resulting random Ising model by exact methods. We implemented the ℓ_1-regularized logistic regression by setting the ℓ_1 penalty as $\lambda_n = \mathcal{O}((\log p)^3 \sqrt{n})$, and solved the convex program using a customized primal-dual algorithm (described in more detail in the full-length version of this paper). We considered various sparsity regimes, including *constant* ($d = \Omega(1)$), *logarithmic* ($d = \alpha \log(p)$), or *linear* ($d = \alpha p$). In each case, we evaluate a given method in terms of its average *precision* (one minus the fraction of falsely included edges), and its *recall* (one minus the fraction of falsely excluded edges). Figure 1 shows results for the case of constant degrees ($d \leq 4$), and graph sizes $p \in \{100, 200, 400\}$, for the AND method (respectively the OR) method, in which an edge (s,t) is included if and only if it is included in the local regressions at both node s *and* (respectively *or*) node t. Note that both the precision and recall tend to one as the number of samples n is increased.

7 Conclusion

We have shown that a technique based on ℓ_1-regularization, in which the neighborhood of any given node is estimated by performing logistic regression subject to an ℓ_1-constraint, can be used for consistent model selection in discrete graphical models. 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. Whereas the current analysis provides sufficient conditions on the triple (n, p, d) that ensure consistent neighborhood selection, it remains to establish necessary conditions as well [11]. Finally, the ideas described here, while specialized in this paper to the binary case, should be more broadly applicable to discrete graphical models.

Acknowledgments

JL and PR supported in part by NSF grants IIS-0427206 and CCF-0625879; MW supported in part by NSF grant DMS-0605165.

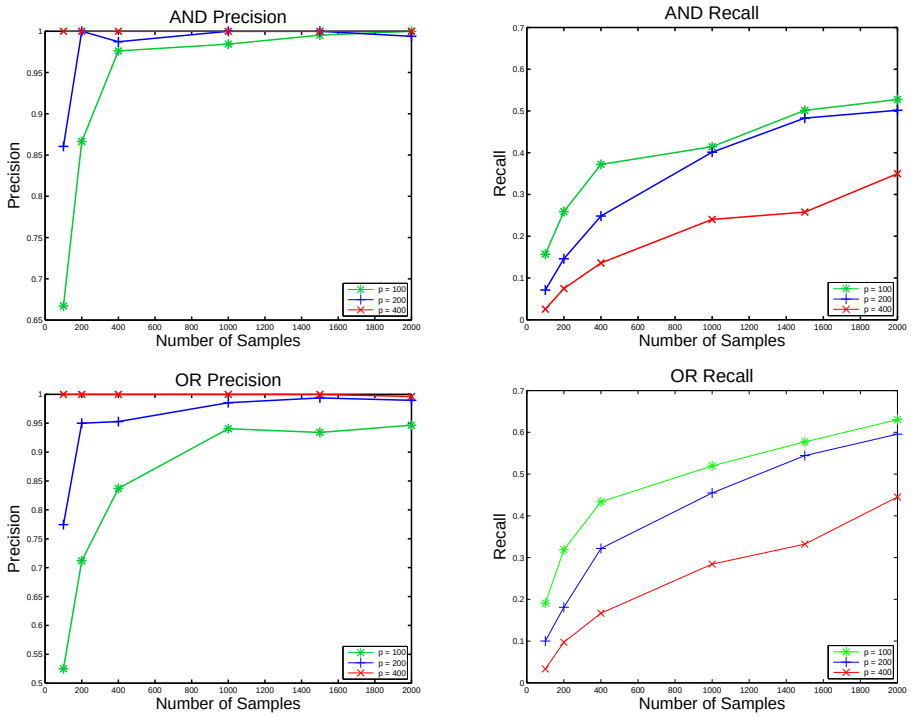

Figure 1. Precision/recall plots using the AND method (top), and the OR method (bottom). Each panel shows precision/recall versus n, for graph sizes $p \in \{100, 200, 400\}$.

References

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

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

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

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

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

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

[7] D. Pollard. *Convergence of stochastic processes*. Springer-Verlag, New York, 1984.

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

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

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

[11] 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.

[12] P. Zhao and B. Yu. Model selection with the lasso. Technical report, UC Berkeley, Department of Statistics, March 2006. Accepted to Journal of Machine Learning Research.