On Graphical Models via Univariate Exponential Family Distributions

Eunho Yang\textsuperscript{1}, Pradeep Ravikumar\textsuperscript{1}, Genevera I. Allen\textsuperscript{2,3}, and Zhandong Liu\textsuperscript{3}

\textsuperscript{1}Department of Computer Science, University of Texas Austin,
\textsuperscript{2} Department of Statistics, Rice University,
\textsuperscript{3} Department of Pediatrics-Neurology, Baylor College of Medicine
& Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital.

Abstract

Undirected graphical models, or Markov networks, are a popular class of statistical models, used in a wide variety of applications. Popular instances of this class include Gaussian graphical models and Ising models. In many settings, however, it might not be clear which subclass of graphical models to use, particularly for non-Gaussian and non-categorical data. In this paper, we consider a general sub-class of graphical models where the node-wise conditional distributions arise from exponential families. This allows us to derive multivariate graphical model distributions from univariate exponential family distributions, such as the Poisson, negative binomial, and exponential distributions. Our key contributions include a class of M-estimators to fit these graphical model distributions; and rigorous statistical analysis showing that these M-estimators recover the true graphical model structure exactly, with high probability. We provide examples of genomic and proteomic networks learned via instances of our class of graphical models derived from Poisson and exponential distributions.

1 Introduction

Undirected graphical models, also known as Markov random fields, are an important class of statistical models that have been extensively used in a wide variety of domains, including statistical physics, natural language processing, image analysis, and medicine. The key idea in this class of models is to represent the joint distribution as a product of clique-wise compatibility functions. Given an underlying graph, each of these compatibility functions depends only on a subset of variables within any clique of the underlying graph. Popular instances of such graphical models include Ising and Potts models (see references in Wainwright and Jordan (2008) for a varied...
set of applications in computer vision, text analytics, and other areas with discrete variables), as well as Gaussian Markov Random Fields (GMRFs), which are popular in many scientific settings for modeling real-valued data. A key modeling question that arises, however, is: how do we pick the clique-wise compatibility functions, or alternatively, how do we pick the form or sub-class of the graphical model distribution (e.g. Ising or Gaussian MRF)? For the case of discrete random variables, Ising and Potts models are popular choices; but these are not best suited for count-valued variables, where the values taken by any variable could range over the entire set of positive integers. Similarly, in the case of continuous variables, Gaussian Markov Random Fields (GMRFs) are a popular choice; but the distributional assumptions imposed by GMRFs are quite stringent. The marginal distribution of any variable would have to be Gaussian for instance, which might not hold in instances when the random variables characterizing the data are skewed \cite{Liu2009}. More generally, Gaussian random variables have thin tails, which might not capture fat-tailed events and variables. For instance, in the finance domain, the lack of modeling of fat-tailed events and probabilities has been suggested as one of the causes of the 2008 financial crisis \cite{Acemoglu2009}.

To address this modeling question, some have recently proposed non-parametric extensions of graphical models. Some, such as the non-paranormal \cite{Liu2009,Lafferty2012} and copula-based methods \cite{Dobra2011,Liu2012a}, use or learn transforms that Gaussianize the data, and then fit Gaussian MRFs to estimate network structure. Others, use non-parametric approximations, such as rank-based estimators, to the correlation matrix, and then fit a Gaussian MRF \cite{Xue2012,Liu2012b}. More broadly, there could be non-parametric methods that either learn the sufficient statistics functions, or learn transformations of the variables, and then fit standard MRFs over the transformed variables. However, the sample complexity of such classes of non-parametric methods is typically inferior to those that learn parametric models. Alternatively, and specifically for the case of multivariate count data, \cite{Lauritzen1996,Bishop2007} have suggested combinatorial approaches to fitting graphical models, mostly in the context of contingency tables. These approaches, however, are computationally intractable for even moderate numbers of variables.

Interestingly, for the case of univariate data, we have a good understanding of appropriate statistical models to use. In particular, a count-valued random variable can be modeled using a Poisson distribution; call-times, time spent on websites, diffusion processes, and life-cycles can be modeled with an exponential distribution; other skewed variables can be modeled with gamma or chi-squared distributions. Here, we ask if we can extend this modeling toolkit from univariate distributions to multivariate graphical model distributions? Interestingly, recent state of the art methods for learning Ising and Gaussian MRFs \cite{Meinshausen2006,Ravikumar2010,Jalali2011} suggest a natural procedure deriving such multivariate graphical models from univariate distributions. The key idea in these recent methods is to learn the MRF graph structure by estimating node-neighborhoods, or by fit-
ting node-conditional distributions of each node conditioned on the rest of the nodes. Indeed, these node-wise fitting methods have been shown to have strong computational as well as statistical guarantees. Here, we consider the general class of models obtained by the following construction: suppose the node-conditional distributions of each node conditioned on the rest of the nodes follows a univariate exponential family. By the Hammersley-Clifford Theorem (Lauritzen 1996), and some algebra as derived in Besag (1974), these node-conditional distributions entail a global multivariate distribution that (a) factors according cliques defined by the graph obtained from the node-neighborhoods, and (b) has a particular set of compatibility functions specified by the univariate exponential family. The resulting class of MRFs, which we call exponential family MRFs, broadens the class of models available off the shelf, from the standard Ising, indicator-discrete, and Gaussian MRFs.

This, thus, provides a principled approach to model multivariate distributions and network structures among a large number of variables. In particular, the class of exponential family MRFs provide a natural way to “extend” univariate exponential families of distributions to the multivariate case, where for many of distributions, multivariate extensions do not exist in an analytical or computationally tractable form. Potential applications for these exponential family graphical models abound. Networks of call-times, time spent on websites, diffusion processes, and life-cycles can be modeled with exponential graphical models; other skewed multivariate data can be modeled with gamma or chi-squared graphical models; while multivariate count data such as from website visits, user-ratings, crime and disease incident reports, bibliometrics could be modeled via Poisson graphical models. A key motivating application for our research was multivariate count data from next-generation genomic sequencing technologies (Marioni et al. 2008). This technology produces read counts of the number of short RNA fragments that have been mapped back to a particular gene; and measures gene expression with less technical variation than, and is thus rapidly replacing, microarrays (Marioni et al. 2008). Univariate count data is typically modeled using Poisson or negative binomial distributions (Li et al. 2011). As Gaussian graphical models have been traditionally used to understand genomic relationships and estimate regulatory pathways from microarray data, Poisson and negative-binomial graphical models could thus be used to analyze this next-generation sequencing data. Furthermore, there is a proliferation of new technologies to measure high-throughput genomic variation in which the data is not even approximately Gaussian (single nucleotide polymorphisms, copy number, methylation, and micro-RNA and gene expression via next-generation sequencing). For this data, a more general class of high-dimensional graphical models could thus lead to important breakthroughs in understanding genomic relationships and disease networks.

The construction of the class of exponential family graphical models also suggests a natural method for fitting such models: node-wise neighborhood estimation via sparsity constrained node-conditional likelihood maximization. A main contribution of this paper is to provide a sparsistency analysis for the recovery of the underlying graph structure of this broad class of MRFs. We note that the presence of non-
linearities arising from the GLM posed subtle technical issues not present in the linear case (Meinshausen and Bühlmann 2006). Indeed, for the specific cases of logistic, and multinomial respectively, Ravikumar et al. (2010); Jalali et al. (2011) derive such a sparsistency analysis via fairly extensive arguments, but which were tuned to the specific cases; for instance they used the fact that the variables were bounded, and the specific structure of the corresponding GLMs. Here we generalize their analysis to general GLMs, which required a subtler analysis as well as a slightly modified M-estimator. We note this analysis might be of independent interest even outside the context of modeling and recovering graphical models. In recent years, there has been a trend towards unified statistical analyses that provide statistical guarantees for broad classes of models via general theorems (Negahban et al. 2010). Our result is in this vein and provides structure recovery for the class of sparsity constrained generalized linear models. We hope that the techniques we introduce might be of use to address the outstanding question of sparsity constrained M-estimation in its full generality.

There has been related work on the simple idea above of constructing joint distributions via specifying node-conditional distributions. Varin and Vidoni (2005); Varin et al. (2011) propose the class of composite likelihood models where the joint distribution is a function of the conditional distributions of subsets of nodes conditioned on other subsets. Besag (1974) discuss such joint distribution constructions in the context of node-conditional distributions belonging to exponential families, but for special cases of joint distributions such as pairwise models. In this paper, we consider the general case of higher-order graphical models for the joint distributions, and univariate exponential families for the node-conditional distributions. Moreover, a key contribution of the paper is that we provide tractable M-estimators with corresponding high-dimensional statistical guarantees and analysis for learning this class of graphical models even under high-dimensional statistical regimes.

2 Exponential Family Graphical Models

Suppose \( X = (X_1, \ldots, X_p) \) is a random vector, with each variable \( X_i \) taking values in a set \( \mathcal{X} \). Let \( G = (V, E) \) be an undirected graph over \( p \) nodes corresponding to the \( p \) variables \( \{X_i\}_{i=1}^p \). The graphical model over \( X \) corresponding to \( G \) is a set of distributions that satisfy Markov independence assumptions with respect to the graph \( G \) (Lauritzen 1996). By the Hammersley-Clifford theorem (Clifford 1990), any such distribution that is strictly positive over its domain also factors according to the graph in the following way. Let \( \mathcal{C} \) be a set of cliques (fully-connected subgraphs) of the graph \( G \), and let \( \{\phi_c(X_c)\}_{c \in \mathcal{C}} \) be a set of clique-wise sufficient statistics. With this notation, any strictly positive distribution of \( X \) within the graphical model family represented by the graph \( G \) takes the form:

\[
P(X) \propto \exp\left\{ \sum_{c \in \mathcal{C}} \theta_c \phi_c(X_c) \right\},
\]  

(1)
where \( \{ \theta_c \} \) are weights over the sufficient statistics. An important special case is a pairwise graphical model, where the set of cliques \( C \) consists of the set of nodes \( V \) and the set of edges \( E \), so that

\[
P(X) \propto \exp \left\{ \sum_{s \in V} \theta_s \phi_s(X_s) + \sum_{(s,t) \in E} \theta_{st} \phi_{st}(X_s, X_t) \right\}.
\]  

(2)

As previously discussed, an important question is how to select the form of the graphical model distribution, which under the above parametrization in (1), translates to the question of selecting the class of sufficient statistics, \( \phi \). As discussed in the introduction, it is of particular interest to derive such a graphical model distribution as a multivariate extension of specified univariate parametric distributions such as negative binomial, Poisson, and others. We next outline a subclass of graphical models that answer these questions via the simple construction: set the node-conditional distributions of each node conditioned on the rest of the nodes as following a univariate exponential family, and then derive the joint distribution that is consistent with these node-conditional distributions. Then, in Section 3, we will study how to learn the underlying graph structure, or the edge set \( E \), for this general class of “exponential family” graphical models. We provide a natural sparsity-encouraging \( M \)-estimator, and sufficient conditions under which the \( M \)-estimator recovers the graph structure with high probability.

2.1 The Form of Exponential Family Graphical Models

A popular class of univariate distributions is the exponential family, whose distribution for a random variable \( Z \) is given by

\[
P(Z) = \exp(\theta B(Z) + C(Z) - D(\theta)),
\]

(3)

with sufficient statistics \( B(Z) \), base measure \( C(Z) \), and log-normalization constant \( D(\theta) \). Such exponential family distributions include a wide variety of commonly used distributions such as Gaussian, Bernoulli, multinomial, Poisson, exponential, gamma, chi-squared, beta, and many others; any of which can be instantiated with particular choices of the functions \( B(\cdot) \), and \( C(\cdot) \). Such exponential family distributions are thus used to model a wide variety of data types including skewed continuous data and count data. Here, we ask if we can leverage this ability to model univariate data to also model the multivariate case. Let \( X = (X_1, X_2, \ldots, X_p) \) be a \( p \)-dimensional random vector; and let \( G = (V, E) \) be an undirected graph over \( p \) nodes corresponding to the \( p \) variables. Could we then derive a graphical model distribution over \( X \) with underlying graph \( G \), from a particular choice of univariate exponential family distribution (3) above?

Consider the following construction. Set the distribution of \( X_s \) given the rest of nodes \( X_{V \setminus s} \) to be given by the above univariate exponential family distribution (3), and where the canonical exponential family parameter \( \theta \) is set to a linear combination
of $k$-th order products of univariate functions \{\(B(X_t)\)\}_{t \in N(s)}$, where \(N(s)\) is the set of neighbors of node \(s\) according to graph \(G\). This gives the following conditional distribution:

\[
P(X_s | X_{V \setminus s}) = \exp \left\{ B(X_s) \left( \theta_s + \sum_{t \in N(s)} \theta_{st} B(X_t) + \sum_{t_2, t_3 \in N(s)} \theta_{s t_2 t_3} B(X_{t_2}) B(X_{t_3}) \right) \right. \\
+ \sum_{t_2, \ldots, t_k \in N(s)} \theta_{s t_2 \ldots t_k} \prod_{j=2}^{k} B(X_{t_j}) \right\} + C(X_s) - \tilde{D}(X_{V \setminus s}),
\]

where \(C(X_s)\) is specified by the exponential family, and \(\tilde{D}(X_{V \setminus s})\) is the log-normalization constant.

By the Hammersley-Clifford theorem, and some elementary calculation, this conditional distribution can be shown to specify the following unique joint distribution \(P(X_1, \ldots, X_p)\):

**Proposition 1.** Suppose \(X = (X_1, X_2, \ldots, X_p)\) is a \(p\)-dimensional random vector, and its node-conditional distributions are specified by (4) given an undirected graph \(G\). Then its joint distribution \(P(X_1, \ldots, X_p)\) belongs to the graphical model represented by \(G\), and is given by:

\[
P(X) = \exp \left\{ \sum_s \theta_s B(X_s) + \sum_{s \in V} \sum_{t \in N(s)} \theta_{st} B(X_s) B(X_t) \right. \\
+ \sum_{s \in V} \sum_{t_2, \ldots, t_k \in N(s)} \theta_{s t_2 \ldots t_k} B(X_s) \prod_{j=2}^{k} B(X_{t_j}) \right\} + \sum_s C(X_s) - A(\theta),
\]

where \(A(\theta)\) is the log-normalization constant.

Proposition 1 thus, provides an answer to our earlier question on selecting the form of a graphical model distribution given a univariate exponential family distribution. When the node-conditional distributions follow a univariate exponential family as in (4), there exists a unique graphical model distribution as specified by (5). One question that remains, however, is whether the above construction, beginning with (4), is the most general possible. In particular, note that the canonical parameter of the node-conditional distribution in (4) is a tensor factorization of the univariate sufficient statistic, which seems a bit stringent. Interestingly, by extending the argument from [Besag 1974], which considers the special pairwise case, and the Hammersley-Clifford Theorem, we can show that indeed (4) and (5) have the most general form.

**Theorem 1.** Suppose \(X = (X_1, X_2, \ldots, X_p)\) is a \(p\)-dimensional random vector, and its node-conditional distributions are specified by an exponential family,

\[
P(X_s | X_{V \setminus s}) = \exp \{ E(X_{V \setminus s}) B(X_s) + C(X_s) - \tilde{D}(X_{V \setminus s}) \},
\]
where the function \( E(X_{V \setminus s}) \), the canonical parameter of exponential family, only depends on variables \( X_t \) in \( N(s) \) (and hence the log-normalization constant \( D(X_{V \setminus s}) \)), and where \( N(s) \) is the set of neighbors of node \( s \) according to an undirected graph \( G = (V, E) \). Further, suppose the corresponding joint distribution factors according to the graph \( G \), with the factors over cliques of size at most \( k \). Then, the conditional distribution in (6) necessarily has the tensor-factorized form in (4), and the corresponding joint distribution has the form in (5).

Theorem 1 thus tells us that under the general assumptions that:

(a) the joint distribution is a graphical model that factors according to a graph \( G \), and has clique-factors of size at most \( k \), and

(b) its node-conditional distribution follows an exponential family,

it necessarily follows that the conditional and joint distributions are given by (4) and (5) respectively.

An important special case is when the joint graphical model distribution has clique factors of size at most two. From Theorem 1, the conditional distribution is given by:

\[
P(X_s | X_{V \setminus s}) = \exp \left\{ \theta_s B(X_s) + \sum_{t \in N(s)} \theta_{st} B(X_s) B(X_t) + C(X_s) - \bar{D}(X_{V \setminus s}) \right\},
\]

(7)

while the joint distribution is given as:

\[
P(X) = \exp \left\{ \sum_{s \in V} \theta_s B(X_s) + \sum_{(s,t) \in E} \theta_{st} B(X_s) B(X_t) + \sum_{s \in V} C(X_s) - A(\theta) \right\}.
\]

(8)

Note that when the univariate sufficient statistic function \( B(\cdot) \) is a linear function \( B(X_s) = X_s \), then the conditional distribution in (7) is precisely a generalized linear model (McCullagh and Nelder 1989) in canonical form,

\[
P(X_s | X_{V \setminus s}) = \exp \left\{ \theta_s X_s + \sum_{t \in N(s)} \theta_{st} X_s X_t + C(X_s) - D(X_{V \setminus s}; \theta) \right\},
\]

(9)

where the canonical parameter of GLMs becomes \( \theta_s + \sum_{t \in N(s)} \theta_{st} X_t \). At the same time, the joint distribution has the form

\[
P(X) = \exp \left\{ \sum_{s \in V} \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t + \sum_{s \in V} C(X_s) - A(\theta) \right\},
\]

(10)

where the log partition function \( A(\cdot) \) in this case is defined as

\[
A(\theta) := \log \int_{X^p} \exp \left\{ \sum_{s \in V} \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t + \sum_{s \in V} C(X_s) \right\} \nu(dx),
\]

(11)
where \( \nu \) is an underlying measure with respect to which the density in (10) is taken.

We will now provide some examples of our general class of “exponential family” graphical model distributions, focusing on the case in (10) with linear functions \( B(X_s) = X_s \). For each of these examples, we will also detail the domain, \( \Theta := \{ \theta : A(\theta) < +\infty \} \), of valid parameters that ensure that the density is normalizable. Indeed, such constraints on valid parameters are typically necessary for the distributions over countable discrete or continuous valued variables.

**Gaussian Graphical Models.** The popular Gaussian graphical model (Speed and Kiiveri 1986) can be derived as an instance of the construction in Theorem 1, with the univariate Gaussian distribution as the exponential family distribution. The univariate Gaussian distribution with known variance \( \sigma^2 \) is given by,

\[
P(Z) \propto \exp \left\{ \frac{\mu Z}{\sigma} - \frac{Z^2}{2\sigma^2} \right\},
\]

where \( Z \in \mathbb{R} \), so that it can be seen to be an exponential family distribution of the form (3), with sufficient statistic \( B(Z) = \frac{Z}{\sigma} \), and base measure \( C(Z) = -\frac{Z^2}{2\sigma^2} \).

Substituting these in (10), we get the distribution,

\[
P(X; \theta^*) \propto \exp \left\{ \frac{1}{\sigma^2} \left( \sum_{s \in V} \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t - \frac{1}{2} \sum_{s \in V} X_s^2 \right) \right\},
\]

which can be seen to be the multivariate Gaussian distribution.

**Ising Models** The Ising model (Wainwright and Jordan 2008) in turn can be derived from the construction in Theorem 1 with the Bernoulli distribution as the univariate exponential family distribution. The Bernoulli distribution is a member of the exponential family of the form (3), with sufficient statistic \( B(Z) = Z \), and base measure \( C(Z) = 0 \).

Substituting these in (10), we get the distribution,

\[
P(X; \theta^*) \propto \exp \left\{ \frac{1}{\sigma^2} \sum_{(s,t) \in E} \theta_{st} X_s X_t - A(\theta) \right\},
\]

where we have ignored the singleton term, i.e. set \( \theta_s = 0 \) for simplicity. The form of the multinomial graphical model, an extension of the Ising model, can also be represented by (10) and has been previously studied in Jalali et al. (2011) and others. The Ising model imposes no constraint on its parameters, \( \{ \theta_{st} \} \), for normalizability, since there are finitely many configurations of the binary random vector \( X \).
Poisson Graphical Models  Poisson graphical models are an interesting instance with the Poisson distribution as the univariate exponential family distribution. The Poisson distribution is a member of the exponential family of the form (3), with sufficient statistic \( B(X) = X \) and \( C(X) = -\log(X!) \), and with variables taking values in the set \( \mathcal{X} = \{0, 1, 2, \ldots\} \). Substituting these in (10), we get the following Poisson graphical model distribution:

\[
P(X) = \exp \left\{ \sum_{s \in V} (\theta_s X_s - \log(X_s!)) + \sum_{(s,t) \in E} \theta_{st} X_s X_t - A(\theta) \right\}. \tag{14}
\]

For this Poisson family, with some calculation, it can be seen that the normalizability condition, \( A(\theta) < +\infty \), entails \( \theta_{st} \leq 0 \ \forall \ s, t \). In other words, the Poisson graphical model can only capture negative conditional relationships between variables.

Exponential Graphical Models  Another interesting instance uses the exponential distribution as the univariate exponential family distribution, with sufficient statistic \( B(X) = -X \) and \( C(X) = 0 \), and with variables taking values in \( \mathcal{X} = \{0\} \cup \mathbb{R}^+ \). Such exponential distributions are typically used for data describing inter-arrival times between events, among other applications. Substituting these in (10), we get the following exponential graphical model distribution:

\[
P(X) = \exp \left\{ -\sum_{s \in V} \theta_s X_s - \sum_{(s,t) \in E^*} \theta_{st} X_s X_t - A(\theta) \right\}. \tag{15}
\]

To ensure that the distribution is valid and normalizable, so that \( A(\theta) < +\infty \), we then require that \( \theta_s > 0, \theta_{st} \geq 0 \ \forall \ s, t \). Because of the negative sufficient statistic, this implies that the exponential graphical model can only capture negative conditional relationships between variables.

3  Statistical Guarantees

In this section, we study the problem of learning the graph structure of an underlying GLM graphical model, given i.i.d. samples. Specifically, we assume that we are given \( n \) samples \( X^n_i := \{X^{(i)}\}_{i=1}^n \), from a GLM graphical model,

\[
P(X; \theta^*) = \exp \left\{ \sum_{s \in V} \theta_s^* X_s + \sum_{(s,t) \in E^*} \theta_{st}^* X_s X_t + \sum_{s} C(X_s) - A(\theta) \right\}. \tag{16}
\]

The goal in graphical model structure recovery is to recover the edges \( E^* \) of the underlying graph \( G = (V, E^*) \). Following Meinshausen and Bühlmann (2006); Ravikumar et al. (2010); Jalali et al. (2011), we will approach this problem via neighborhood
estimation: where we estimate the neighborhood of each node individually, and then stitch these together to form the global graph estimate. Specifically, if we have an estimate \( \hat{\mathcal{N}}(s) \) for the true neighborhood \( \mathcal{N}^*(s) \), then we can estimate the overall graph structure as,

\[
\hat{E} = \bigcup_{s \in V} \bigcup_{t \in \hat{\mathcal{N}}(s)} \{ (s,t) \}.
\]  

(17)

The problem of graph structure recovery can thus be reduced to the problem of recovering the neighborhoods of all the nodes in the graph. In order to estimate the neighborhood of any node in turn, we consider the sparsity constrained conditional MLE. Note that given the joint distribution in (16), the conditional distribution of \( X_s \) given the rest of the nodes is given by,

\[
P(X_s|X_{V\setminus s}) = \exp \left\{ X_s \left( \theta^*_s + \sum_{t \in N(s)} \theta^*_{st} X_t \right) + C(X_s) - D\left( \theta^*_s + \sum_{t \in N(s)} \theta^*_{st} X_t \right) \right\}.
\]  

(18)

Let \( \theta^*(s) \) be a set of parameters related to the node-conditional distribution of node \( X_s \), i.e. \( \theta^*(s) = \{ \theta^*_s, \theta^*_{st} \} \in \mathbb{R} \times \mathbb{R}^{p-1} \) where \( \theta^*_s \) is a zero-padded vector, with entries \( \theta^*_{st} \) for \( t \in N(s) \) and \( \theta^*_{st} = 0 \), for \( t \not\in N(s) \). Given \( n \) samples \( X^n_s = \{ X^{(i)}_s \}_{i=1}^n \), we can solve the \( \ell_1 \) regularized conditional log-likelihood loss for each node \( X_s \):

\[
\min_{\theta(s) \in \mathbb{R} \times \mathbb{R}^{p-1}} \ell(\theta(s); X^n_s) + \lambda_n \| \theta_s \|_1,
\]  

(19)

where \( \ell(\theta(s); X^n_s) \) is the conditional log-likelihood of the distribution (18),

\[
\ell(\theta(s); X^n_s) := -\frac{1}{n} \log \prod_{i=1}^n P(X^{(i)}_s|X_{V\setminus s}, \theta(s)) \\
= -\frac{1}{n} \sum_{i=1}^n \left\{ -X^{(i)}_s \left( \theta_s + \langle \theta_{\setminus s}, X^{(i)}_{V\setminus s} \rangle \right) + D\left( \theta_s + \langle \theta_{\setminus s}, X^{(i)}_{V\setminus s} \rangle \right) \right\}.
\]

Given the solution \( \hat{\theta}(s) \) of the M-estimation problem above, we then estimate the node-neighborhood of \( s \) as \( \hat{\mathcal{N}}(s) = \{ t \in V\setminus s : \hat{\theta}_{st} \neq 0 \} \). In the sequel when we focus on a fixed node \( s \in V \), we will overload notation, and use \( \theta \in \mathbb{R} \times \mathbb{R}^{p-1} \) as the parameters of the conditional distribution, suppressing dependence on the node \( s \).

3.1 Conditions

A key quantity in the analysis is the Fisher Information matrix, \( Q_* = \nabla^2 \ell(\theta^*(s); X^n_s) \), which is the Hessian of the node-conditional log-likelihood. In the following, we again will simply use \( Q_0 \) instead of \( Q_* \), where the reference node \( s \) should be understood implicitly. We also use \( S = \{ (s,t) : t \in N(s) \} \) to denote the true neighborhood of
node $s$, and $S^c$ to denote its complement. We use $Q^*_{SS}$ to denote the $d \times d$ sub-matrix indexed by $S$ where $d$ is the maximum node degree. Our first two conditions, mirroring those in Ravikumar et al. (2010), are as follows.

**Condition 1** (Dependency condition). There exists a constant $\lambda_{\min} > 0$ such that $\lambda_{\min}(Q^*_{SS}) \geq \lambda_{\min}$ so that the sub-matrix of Fisher Information matrix corresponding to true neighborhood has bounded eigenvalues. Moreover, there exists a constant $\lambda_{\max} < \infty$ such that $\lambda_{\max}(\mathbb{E}[X_s X^T_s]) \leq \lambda_{\max}$.

These condition can be understood as ensuring that variables do not become overly dependent. We will also need an incoherence or irrepresentable condition on the Fisher information matrix as in Ravikumar et al. (2010).

**Condition 2** (Incoherence condition). There exists a constant $\alpha > 0$, such that $\max_{t \in S^c} \|Q^*_{tS}(Q^*_{SS})^{-1}\|_1 \leq 1 - \alpha$.

This condition, standard in high-dimensional analyses, can be understood as ensuring that irrelevant variables do not exert an overly strong effect on the true neighboring variables.

A key technical facet of the linear, logistic, and multinomial models in Meinshausen and Bühlmann (2006); Ravikumar et al. (2010); Jalali et al. (2011), used heavily in their proofs, was that the random variables $\{X_s\}$ there were bounded with high probability. Unfortunately, in the general GLM distribution in (18), we cannot assume this explicitly. Nonetheless, we show that we can analyze the corresponding regularized M-estimation problems under the following mild conditions on the log-partition functions of the joint and node-conditional distributions.

**Condition 3.** The log-partition function $A(\cdot)$ of the joint distribution (16) satisfies: For all $s \in V$, (i) there exist constants $\kappa_m, \kappa_v$ such that the first and the second moment satisfy $\mathbb{E}[X_s] \leq \kappa_m$ and $\mathbb{E}[X_s^2] \leq \kappa_v$, respectively. Additionally, we have a constant $\kappa_h$ for which $\max_{u:|u|\leq 1} \frac{\partial^2 A(\theta)}{\partial \eta^2}(\{\theta_s^* + u, \theta^*\}) \leq \kappa_h$, and (ii) for scalar variable $\eta$, we define a function which is slightly different from (11):

$$\tilde{A}_s(\eta; \theta) := \log \int_{\mathcal{X}_p} \exp \left\{ \eta X_s^2 + \sum_{s \in V} \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t + \sum_{s \in V} C(X_s) \right\} \nu(dx),$$

where $\nu$ is an underlying measure with respect to which the density is taken. Then, there exists a constant $\kappa_h$ such that $\max_{u:|u|\leq 1} \frac{\partial^2 \tilde{A}_s(\eta; \theta)}{\partial \eta^2}(u) \leq \kappa_h$.

**Condition 4.** For all $s \in V$, the log-partition function $D(\cdot)$ of the node-wise conditional distribution (18) satisfies: there exist functions $\kappa_1(n, p)$ and $\kappa_2(n, p)$ (that depend on the exponential family) such that, for all feasible pairs of $\theta$ and $X$, $|D^y(a)| \leq \kappa_1(n, p)$ where $a \in [b, b + 4\kappa_2(n, p) \max\{\log n, \log p\}]$ for $b := \theta_s + \langle \theta_s, X_s \rangle$. Additionally, $|D^\eta(b)| \leq \kappa_3(n, p)$ for all feasible pairs of $\theta$ and $X$. Note that $\kappa_1(n, p), \kappa_2(n, p)$ and $\kappa_3(n, p)$ are functions that might be dependent on $n$ and $p$, which affect our main theorem below.
3.2 Statement of the Sparsistency Result

Armed with the conditions above, we can show that the random vectors $X$ following the GLM graphical model distribution in (16) are suitably well-behaved:

**Proposition 2.** Suppose $X$ is a random vector with the distribution specified in (16). Then, for $\delta \leq \min\{2\kappa_v/3, \kappa_h + \kappa_v\}$, and some constant $c > 0$,

$$P \left( \frac{1}{n} \sum_{i=1}^{n} (X_s^{(i)})^2 \geq \delta \right) \leq \exp \left( -cn \delta^2 \right),$$

**Proposition 3.** Suppose $X$ is a random vector with the distribution specified in (16). Then, for any positive constant $\delta$, and some constant $c > 0$,

$$P \left( |X_s| \geq \delta \log \eta \right) \leq c\eta^{-\delta}.$$

These propositions are key to the following sparsistency result for the general family of GLM graphical model distributions (16).

**Theorem 2.** Consider a GLM graphical model distribution as specified in (16), with true parameter $\theta^*$ and associated edge set $E^*$ that satisfies Conditions 3 and 4. Suppose that $\min_{(s,t) \in E^*} |\theta^*_{st}| \geq \frac{10}{\alpha} \sqrt{d\lambda_n}$, where $d$ is the maximum neighborhood size. Suppose also that the regularization parameter is chosen such that $M_1 \frac{(2-\alpha)}{\alpha} \kappa_1(n,p) \sqrt{\log p \over n} \leq \lambda_n \leq M_2 \frac{(2-\alpha)}{\alpha} \kappa_1(n,p) \kappa_2(n,p)$ for some constants $M_1, M_2 > 0$. Then, there exist positive constants $L, c_1, c_2$ and $c_3$ such that if $n \geq Ld^2 \kappa_1(n,p) \kappa_3(n,p) \log p \{\max\{\log n, \log p\}\}^2$, then with probability at least $1 - c_1 \max\{n, p\}^{-2} - \exp(-c_2n) - \exp(-c_3n)$, the following statements hold.

(a) **(Unique Solution)** For each node $s \in V$, the solution of the M-estimation problem in (19) is unique, and

(b) **(Correct Neighborhood Recovery)** The M-estimate also recovers the true neighborhood exactly, so that $\hat{N}(s) = N(s)$.

Note that if the neighborhood of each node is recovered with high probability, then by a simple union bound, the estimate in (17), $\hat{E} = \cup_{s \in V} \cup_{t \in \hat{N}(s)} \{(s, t)\}$ is equal to the true edge set $E^*$ with high-probability.
Remark. It is instructive to compare the sample complexity derived in Theorem 2 to that derived for multinomial graphical models in Jalali et al. (2011), where each variable takes values in a finite set $X$ of cardinality $m$, and where the statements of the theorem is shown to hold with sample size $n \geq Ld^2 \log p(m-1)^2$. In our theorem that applies to general GLMs, we cannot of course assume that each variable $X$ is bounded by $m-1$. Instead, in the proof of our theorem, we leverage Propositions 2 and 3 and show that each variable $X$ is bounded by $\max\{\log n, \log p\}$ with high probability; thus matching the sample complexity as derived in Jalali et al. (2011) with $m = \max\{\log n, \log p\}$.

In the following subsections, we investigate the consequences of Theorem 2 for the sparsistency of specific instances of our general GLM graphical model family.

### 3.3 Statistical Guarantees for Gaussian MRFs, Ising Models, Exponential Graphical Models

In order to apply Theorem 2 to a specific instance of our general GLM graphical model family, we need to specify the terms $\kappa_1(n, p), \kappa_2(n, p)$ and $\kappa_3(n, p)$ defined in Condition 4. It turns out that we can specify these terms for the Gaussian graphical models, Ising models and Exponential graphical model distributions, discussed in Section 2, in a similar manner, since the node-conditional log-partition function $D(\cdot)$ for all these distributions can be upper bounded by some constant independent of $n$ and $p$. In particular, we can set $\kappa_1(n, p) := \kappa_1, \kappa_2(n, p) := \infty$ and $\kappa_1(n, p) := \kappa_3$ where $\kappa_1$ and $\kappa_3$ now become some constants depending on the distributions.

**Gaussian MRFs.** Recall that the node conditional distribution for Gaussian MRFs follow a univariate Gaussian distribution:

$$P(X_s|X_{V\setminus s}) \propto \exp \left\{ \frac{1}{\sigma^2} \left( X_s\left( \theta_s + \sum_{t \in N(s)} \theta_{st} X_t \right) - \frac{1}{2} X_s^2 - \frac{1}{2} X_s^2 \right) \right\}.$$

The node-conditional log-partition function $D(\cdot)$ can thus be written as $D(\eta) := -\frac{1}{2} \eta^2$, so that $|D''(\eta)| = 1$ and $D'''(\eta) = 0$. We can thus set $\kappa_1 = 1$ and $\kappa_3 = 0$.

**Ising Models.** Similarly for Ising models, node conditional distribution follows a Bernoulli distribution:

$$P(X_s|X_{V\setminus s}) = \exp \left\{ X_s\left( \sum_{t \in N(s)} \theta_{st} X_t \right) - \log \left( 1 + \exp \left( \sum_{t \in N(s)} \theta_{st} X_t \right) \right) \right\}.$$

The node-conditional log-partition function $D(\cdot)$ can thus be written as $D(\eta) := \log \left( 1+\exp \eta \right)$, so that for any $\eta$, $|D''(\eta)| = \frac{\exp(\eta)}{(1+\exp(\eta))^2} \leq \frac{1}{4}$ and $|D'''(\eta)| = \frac{\exp(\eta)(1-\exp(\eta))}{(1+\exp(\eta))^3} < \frac{1}{4}$. Hence, we can set $\kappa_1 = \kappa_3 = 1/4$. 

13
Exponential Graphical Models. Lastly, for exponential graphical models, we have

\[ P(X_s | X_{V\setminus s}) = \exp \left\{ -X_s\left( \theta_s + \sum_{t \in N(s)} \theta_{st}X_t \right) - \log \left( \theta_s + \sum_{t \in N(s)} \theta_{st}X_t \right) \right\}. \]

The node-conditional log-partition function \( D(\cdot) \) can thus be written as \( D(\eta) := \log \eta, \) with \( \eta = \theta_s + \sum_{t \in N(s)} \theta_{st}X_t. \) Recall from Section 2.1 that the node parameters are strictly positive \( \theta_s > 0, \) and the edge-parameters are positive as well, \( \theta_{st} \geq 0, \) as are the variables themselves \( X_t \geq 0. \) Thus, under the additional constraint that \( \theta_s > a_0 \) where \( a_0 \) is a constant smaller than \( \theta^*_s, \) we have that \( \eta := \theta_s + \sum_{t \in N(s)} \theta_{st}X_t \geq a_0. \) Consequently, \( |D''(\eta)| = \frac{1}{\eta^2} \leq \frac{1}{a_0^2} \) and \( |D'''(\eta)| = \frac{2}{\eta^3} \leq \frac{2}{a_0^3}. \) We can thus set \( \kappa_1 = \frac{1}{a_0^2} \) and \( \kappa_3 = \frac{2}{a_0^3}. \)

Armed with these derivations, we recover the following result on the sparsistency of Gaussian, Ising and Exponential graphical models, as a corollary of Theorem 2:

**Corollary 1.** Consider a Gaussian MRF (12) or Ising model (13) or Exponential graphical model (15) distribution with true parameter \( \theta^*, \) and associated edge set \( E^*, \) and which satisfies Conditions 1-3. Suppose that \( \min_{(s,t) \in E^*} |\theta^*_{st}| \geq \frac{10}{\lambda_{\min}} \sqrt{d \log n}. \) Suppose also that the regularization parameter is set so that \( M \left( \frac{2-\alpha}{\alpha} \right) \sqrt{\kappa_1} \sqrt{\frac{\log p}{n}} \leq \lambda_n \) for some constant \( M > 0. \) Then, there exist positive constants \( L, c_1, c_2 \) and \( c_3 \) such that if \( n \geq L \kappa_1 \kappa_3 d^2 \log p (\max\{\log n, \log p\})^2, \) then with probability at least \( 1 - c_1 \max\{n, p\}^{-2} - \exp(-c_2n) - \exp(-c_3n), \) the statements on the uniqueness of the solution and correct neighborhood recovery, in Theorem 2 hold.

**Remark.** As noted, our models and theorems are quite general, extending well beyond the popular Ising and Gaussian graphical models. The graph structure recovery problem for Gaussian models was studied in Meinshausen and Bühlmann (2006) especially for the regime where the neighborhood sparsity index is sublinear, meaning that \( d/p \to 0. \) Besides the sublinear scaling regime, Corollary 1 can be adapted to entirely different types of scaling, such as the linear regime where \( d/p \to \alpha \) for some \( \alpha > 0 \) (see Wainwright (2009) for details on adaptations to sublinear scaling regimes). Moreover, with \( \kappa_1 \) and \( \kappa_3 \) as defined above, Corollary 1 exactly recovers the result in Ravikumar et al. (2010) for the Ising models as a special case.

### 3.4 Statistical Guarantees for Poisson Graphical Models

We now consider the Poisson graphical model. Again, to derive the corresponding corollary of Theorem 2, we need to specify the terms \( \kappa_1(n, p), \kappa_2(n, p) \) and \( \kappa_3(n, p) \) defined in Condition 4. Recall that the node-conditional distribution of Poisson graphi-
The node-conditional log-partition function $D(\cdot)$ can thus be written as $D(\eta) := \exp \eta$, with $\eta = \theta_s + \sum_{t \in N(s)} \theta_{st} X_t$. Noting that the variables $\{X_t\}$ are range over positive integers, and that feasible parameters $\theta_{st}$ are negative, we obtain

$$D''(\eta) = D''(\theta_s + \langle \theta_s, X_{V\setminus s} \rangle + 4\kappa_2(n, p) \log p') = \exp (\theta_s + \langle \theta_s, X_{V\setminus s} \rangle + 4\kappa_2(n, p) \log p') \leq \exp (\theta_s + 4\kappa_2(n, p) \log p'),$$

where $p' = \max \{n, p\}$. Suppose that we restrict our attention on the subfamily where $\theta_s \leq a_0$ for some positive constant $a_0$. Then, if we choose $\kappa_2(n, p) := 1/(4 \log p')$, we then obtain $\theta_s + 4\kappa_2(n, p) \log p' \leq a_0 + 1$, so that setting $\kappa_1(n, p) := \exp(a_0 + 1)$ would satisfy Condition 11. Similarly, we obtain $D''(\theta_s + \langle \theta_s, X_{V\setminus s} \rangle) = \exp (\theta_s + \langle \theta_s, X_{V\setminus s} \rangle) \leq \exp(a_0 + 1)$, so that we can set $\kappa_3(n, p)$ to $\exp(a_0 + 1)$.

Armed with these settings, we recover the following corollary for Poisson graphical models:

**Corollary 2.** Consider a Poisson graphical model distribution as specified in (14), with true parameters $\theta^*$, and associated edge set $E^*$, that satisfies Conditions 13. Suppose that $\min_{(s,t) \in E^*} |\theta^*_{st}| \geq 10^{-10} \lambda_{\min} \sqrt{d} \lambda_n$. Suppose also that the regularization parameter is chosen such that $M_1(2^{-\alpha}) \sqrt{\kappa_1} \sqrt{\frac{\log p}{n}} \leq \lambda_n \leq M_2 \kappa_1 \frac{2^{-\alpha}}{\max(\log n, \log p)}$ for some constants $M_1, M_2 > 0$. Then, there exist positive constants $L, c_1, c_2$ and $c_3$ such that if $n \geq L d^2 \kappa_1 \kappa_3^2 \log p (\max\{\log n, \log p\})^2$, then with probability at least $1 - c_1 \max\{n, p\}^{-2} - \exp(-c_2 n) - \exp(-c_3 n)$, the statements on the uniqueness of the solution and correct neighborhood recovery, in Theorem 2 hold.

### 3.5 Proof of Theorem 2

In this section, we sketch the proof of Theorem 2 following the primal-dual witness proof technique in [Wainwright (2009); Ravikumar et al. (2010)]. We first note that the optimality condition of the convex program (19) can be written as

$$\nabla \ell(\hat{\theta}; X^n) + \lambda_n \hat{Z} = 0,$$

(21)

where $\hat{Z}_{V\setminus s}$ is an element of the subgradient of $\|\hat{\theta}_{s}\|_1$: $\hat{Z}_{st} = \text{sign}(\hat{\theta}_{st})$ if $\hat{\theta}_{st} \neq 0$, and $|\hat{Z}_{st}| \leq 1$ otherwise; and where we set $\hat{Z}_s = 0$, since the nodewise terms $\{\theta_s\}$ are not penalized in the $M$-estimation problem (19).

Note that in a high-dimensional regime with $p \gg n$, the convex program (19) is not necessarily strictly convex, so that it might have multiple optimal solutions. However, the following lemma, adapted from [Ravikumar et al. (2010)], shows that nonetheless the solutions share their support set under certain conditions:
Lemma 1. Let $S$ be the true support set of the edge parameters, so that $\theta^*_i \neq 0$ iff $(i,j) \in S$. Suppose that there exists an primal optimal solution $\hat{\theta}$ with associated subgradient $\hat{Z}$ s.t. $\|\hat{Z}_S\| < 1$. Then, any optimal solution $\hat{\theta}$ will satisfy $\hat{\theta}_S = 0$.

We use this lemma to prove the theorem following the primal-dual witness proof technique in Wainwright (2009); Ravikumar et al. (2010). Specifically, we explicitly construct a pair $(\hat{\theta}, \hat{Z})$ as follows (denoting the true support set of the edge parameters by $S$):

(a) We set $\hat{\theta}_S = \arg\min_{\theta_S \in \mathbb{R}_+,(\theta_{S,0}) \in \mathbb{R}^{p-1}} \{\ell(\theta; X^n_1) + \lambda_n \|\theta_S\|_1\}$, and $\hat{Z}_S = \text{sign}(\hat{\theta}_S)$.

(b) We set $\hat{\theta}_{Sc} = 0$.

(c) We set $\hat{Z}_{Sc}$ to satisfy the condition (21) with $\hat{\theta}$ and $\hat{Z}_S$.

By construction, the support of $\hat{\theta}$ is included in the true support $S$ of $\theta^*$, so that we would finish the proof of the theorem provided (a) $\hat{\theta}$ satisfies the stationary condition of (19), as well as the condition in Lemma 1; and (b) the support of $\hat{\theta}$ is not strictly within the true support $S$. We term these conditions strict dual feasibility, and sign consistency respectively, and in the sequel, show that they hold with high probability.

It can be seen that from the conditions of Lemma 1 that strict dual feasibility holds if $\|\hat{Z}_S\|_{\infty} < 1$ with high probability. We will now rewrite the sub-gradient optimality condition (21) as

$$\nabla^2 \ell(\theta^*; X^n_1)(\hat{\theta} - \theta^*) = -\lambda_n \hat{Z} + W^n + R^n,$$

where $W^n := -\nabla \ell(\theta^*; X^n_1)$ is the sample score function (that we will show is small with high probability), and $R^n$ is the remainder term after coordinate-wise applications of the mean value theorem; $R^n_i = \left[\nabla^2 \ell(\theta^*; X^n) - \nabla^2 \ell(\theta^{(j)}; X^n)\right]_{ij}(\hat{\theta} - \theta^*)$, for some $\theta^{(j)}$ on the line between $\hat{\theta}$ and $\theta^*$, and with $[\cdot]_j^T$ being the $j$-th row of a matrix.

Recalling the notation for the Fisher Information matrix $Q^* := \nabla^2 \ell(\theta^*; X^n)$, we then have

$$Q^*(\hat{\theta} - \theta^*) = -\lambda_n \hat{Z} + W^n + R^n.$$

In the sequel, we provide lemmas that respectively control various terms in the above expression: the score term $W^n$, the deviation $\hat{\theta} - \theta^*$, and the remainder term $R^n$. The first lemma controls the score term $W^n$.

Lemma 2. Suppose that we set $\lambda_n$ to satisfy $\frac{8(2-\alpha)}{\alpha} \kappa_1(n,p)\kappa_4 \sqrt{\log p} \leq \lambda_n \leq \frac{4(2-\alpha)}{\alpha} \kappa_1(n,p)\kappa_2(n,p)\kappa_4$ for some constant $\kappa_4 \leq \min\{2\kappa_v/3, 2\kappa_h + \kappa_v\}$. Suppose also that $n \geq \frac{8\kappa_2^2}{\kappa_4^2} \log p$. Then, given the mutual incoherence parameter $\alpha \in (0, 1]$,

$$P\left(\frac{2 - \alpha}{\lambda_n} \|W^n\|_{\infty} \leq \frac{\alpha}{4}\right) \geq 1 - c_1 p'^{-2} - \exp(-c_2 n) - \exp(-c_3 n),$$

where $p' := \max\{n, p\}$.  

16
The next lemma controls the deviation $\hat{\theta} - \theta^*$:

**Lemma 3.** Suppose that $\lambda_n d \leq \frac{\lambda_{\min}^2}{100\lambda_{\max} \kappa_3(n,p) \log p'}$ and $\|W^n\|_{\infty} \leq \frac{\lambda_n}{3}$. Then, we have

$$P\left( \|\hat{\theta}_S - \theta^*_S\|_2 \leq \frac{5}{\lambda_{\min}} \sqrt{d\lambda_n} \right) \geq 1 - c_1 p'^{-2}. \quad (22)$$

for some constant $c_1 > 0$.

The last lemma controls the Taylor series remainder term $R^n$:

**Lemma 4.** If $\lambda_n d \leq \frac{\lambda_{\min}^2}{400\lambda_{\max} \kappa_3(n,p) \log p'} 2^{-\alpha}$, and $\|W^n\|_{\infty} \leq \frac{\lambda_n}{4}$, then we have

$$P\left( \frac{\|R^n\|_{\infty}}{\lambda_n} \leq \frac{\alpha}{4(2 - \alpha)} \right) \geq 1 - c_1 p'^{-2}. \quad (23)$$

for some constant $c_1 > 0$.

The proof then follows from Lemmas 2, 3 and 4 in a straightforward fashion, following Ravikumar et al. (2010). Consider the choice of regularization parameter $\lambda_n = \frac{8(2 - \alpha)}{\alpha} \sqrt{\kappa_1(n,p) \kappa_4} \sqrt{\frac{\log p}{n}}$. For a sample size greater $n \geq \max\{\kappa_1(n,p) \kappa_4, 8\kappa_2^2\}$ log $p$, the conditions of Lemma 2 are satisfied, so that we may conclude that $\|W^n\|_{\infty} \leq \frac{\lambda_n}{4}$ with high probability. Moreover, with a sufficiently large sample size such that $n \geq L' \left(\frac{2 - \alpha}{\alpha}\right)^4 d^2 \kappa_1(n,p) \kappa_3(n,p)^2 \log p (\log p')^2$ for some constant $L' > 0$, the conditions of Lemma 3 and 4 in turn are satisfied, and hence the resulting statements (22) and (23) of Lemmas 3 and 4 hold with high probability.

**Strict dual feasibility.** Following Ravikumar et al. (2010), we obtain

$$\left\| Z_{S^c} \right\|_{\infty} \leq \left\| Q_{S^c}^* (Q_{S^c} S)^{-1} \right\|_{\infty} \left[ \frac{\|W^n_S\|_{\infty}}{\lambda_n} + \frac{\|R^n_S\|_{\infty}}{\lambda_n} + 1 \right] + \left[ \frac{\|W^n_S\|_{\infty}}{\lambda_n} + \frac{\|R^n_S\|_{\infty}}{\lambda_n} \right] \leq (1 - \alpha) + (2 - \alpha) \left[ \frac{\|W^n\|_{\infty}}{\lambda_n} + \frac{\|R^n\|_{\infty}}{\lambda_n} \right] \leq (1 - \alpha) + \frac{\alpha}{4} + \frac{\alpha}{4} = 1 - \frac{\alpha}{2} < 1.$$ 

**Correct sign recovery.** To guarantee that the support of $\hat{\theta}$ is not strictly within the true support $S$, it suffices to show that $\|\hat{\theta}_S - \theta^*_S\|_{\infty} \leq \frac{\theta_{\min}^*}{2}$. From Lemma 3 we have $\|\hat{\theta}_S - \theta^*_S\|_{\infty} \leq \|\hat{\theta}_S - \theta^*_S\|_2 \leq \frac{5}{\lambda_{\min}} \sqrt{d\lambda_n} \leq \frac{\theta_{\min}^*}{2}$ as long as $\theta_{\min}^* \geq \frac{10}{\lambda_{\min}} \sqrt{d\lambda_n}$. This completes the proof.

## 4 Experiments

### 4.1 Simulation Studies

We provide a small simulation study that corroborates our sparsistency results; specifically Corollary 4 for the exponential graphical model, where node conditional distributions follow an exponential distribution, and Corollary 2 for the Poisson graphical
model, where node conditional distributions follow a Poisson distribution. We instantiated the corresponding exponential and Poisson graphical model distributions in (15) and (14) for lattice (4 nearest neighbor) graphs, with varying number of nodes, \( p \in \{64, 100, 169, 225\} \), and with identical edge weights for all edges. We generated i.i.d. samples from these distributions using Gibbs sampling. We repeated each simulation 50 times and measured the empirical probability over the 50 trials that our penalized graph estimate in (17) successfully recovered all edges. We solved our sparsity-constrained \( M \)-estimation problem by setting \( \lambda_n \) to a constant factor of \( \sqrt{\frac{\log p}{n}} \), following our corollaries. The left panels of Figure 1(a) and Figure 1(b) show the empirical probability of successful edge recovery. In the right panel, we plot the empirical probability against a re-scaled sample size \( \beta = n/(c \log p) \) for some constant \( c \). According to our corollaries, this rescaled sample size, as a function of \( n \) and \( p \), drives the probability of success in recovering the graph structure. Thus, we would expect the empirical curves for different problem sizes to align with this re-scaled sample-size on the horizontal axis, a result clearly seen in the right panels of Figure 1. This small numerical study thus corroborates our theoretical sparsistency results.

4.2 Real Data Examples

To demonstrate the versatility of our family of graphical models, we also provide two real data examples: (a) a meta-miRNA inhibitory network estimated by the Poisson graphical model, Figure 2 (left); and (b) an inhibitory cell signaling network estimated by the exponential graphical model, Figure 2 (right).

4.2.1 Poisson Graphical Model: Meta-miRNA Inhibitory Network

Gaussian graphical models have often been used to study high-throughput genomic networks learned from microarray data (Pe’er et al. 2001; Friedman 2004; Wei and Li 2007). Many high-throughput technologies, however, do not produce even approximately Gaussian data, so that our class of graphical models could be particularly important for estimating genomic networks from such data. We demonstrate the applicability of our class of models by learning a meta-miRNA inhibitory network for breast cancer estimated by a Poisson graphical model. Level III breast cancer miRNA expression (Cancer Genome Atlas Research Network 2012) as measured by next generation sequencing was downloaded from the TCGA portal (http://tcga-data.nci.nih.gov/tcga/). MicroRNAs (miRNA) are short RNA fragments that are thought to be post-transcriptional regulators, enhancing or inhibiting translation. Measuring miRNA expression by high-throughput sequencing results in count data that is zero-inflated, highly skewed, and whose total count volume depends on experimental conditions (Li et al. 2011). Data was processed to be approximately Poisson by following the steps in (Allen and Liu 2012). In brief, the data was quantile corrected to adjust for sequencing depth (Bullard et al. 2010); the miRNAs with little variation across the samples, the bottom 50%, were filtered out; and the data was
Figure 1: Probabilities of successful support recovery for the (a) exponential MRF, grid structure with parameters $\theta_s^* = 0.1$ and $\theta_{st}^* = 1$, and the (b) Poisson MRF, grid structure with parameters $\theta_s^* = 2$ and $\theta_{st}^* = -0.1$. The empirical probability of successful edge recovery over 50 replicates is shown versus the sample size $n$ (left), and versus the re-scaled sample size $\beta = n/(c \log p)$ (right). The empirical curves align for the latter, thus verifying the rates obtained in our sparsistency analysis.

adjusted for possible over-dispersion using a power transform and a goodness of fit test (Li et al. 2011; Allen and Liu 2012). We also tested for batch effects in the resulting data matrix consisting of 544 subjects and 262 miRNAs: we fit a Poisson ANOVA model (Leek et al. 2010), and only found 4% of miRNAs to be associated with batch labels; and thus no significant batch association was detected. As Poisson graphical models are restricted to capture negative conditional relationships, or inhibitory effects, we first pre-processed our data to form clusters of positively correlated miRNAs or “meta-miRNAs”. Specifically, we used hierarchical clustering with average linkage to cluster the miRNAs into tightly positively correlated groups, 32 in total. The mediod, or median centroid of each cluster, was then taken to be the driver miRNA, and formed the nodes of our meta-miRNA network.

A Poisson graphical model was fit to the meta-miRNA data by performing neigh-
borhood selection with the sparsity of the graph determined by stability selection (Liu et al. 2010). Our results in Figure 2 (left) are consistent with the cancer genomic literature. First, the meta-miRNA inhibitory network has three major hubs. Two of these, mir-519 and mir-520, are known to be breast cancer tumor suppressors, suppressing growth by reducing HuR levels (Abdelmohsen et al. 2010) and by targeting NF-KB and TGF-beta pathways (Keklikoglou et al. 2011) respectively. The third major hub, mir-3156, is a miRNA of unknown function; from its major role in our network, we hypothesize that mir-3156 is also associated with tumor suppression. Also interestingly, let-7, a well-known miRNA involved in tumor metastasis (Yu et al. 2007), plays a central role in our network, sharing edges with the five largest hubs. This suggests that our Poisson graphical model has recovered relevant negative relationships between miRNAs with the five major hubs acting as suppressors, and

Figure 2: Meta-miRNA inhibitory network for breast cancer learned via Poisson graphical models (left) and a inhibitory cell signaling network learned from flow cytometry data via exponential graphical models (right). For the meta-miRNA inhibitory network, miRNA-sequencing data from TCGA was processed into tightly correlated clusters, meta-miRNAs, with the driver miRNAs identified for each cluster taken as the set of nodes for our network. The Poisson network reveals major inhibitory relationships between three hub miRNAs, two of which have been previously identified as tumor suppressors in breast cancer. For the cell signaling network, an exponential graphical model was fit to un-transformed flow cytometry data measuring 11 proteins. The exponential network identifies PKA (protein kinase A) as a major inhibitor, consistent with previous results.
the central let-7 miRNA and those connected to each of the major hubs acting as enhancers of tumor progression in breast cancer.

4.2.2 Exponential Graphical Model: Inhibitory Cell-Signaling Network

We demonstrate our exponential graphical model, derived from the univariate exponential distribution, using a protein signaling example [Sachs et al. 2005]. Multifluorescent flow cytometry was used to measure the presence of eleven proteins \((p = 11)\) in \(n = 7462\) cells. This data set was first analyzed using Bayesian Networks in Sachs et al. (2005) and then using the graphical lasso algorithm in Friedman et al. (2007). Measurements from flow-cytometry data typically follow a left skewed distribution. Thus to model such data, these measurements are typically normalized to be approximately Gaussian using a log transform after shifting the data to be non-negative [Herzenberg et al. 2006]. Here, we demonstrate the applicability of our exponential graphical models to recover networks directly from continuous skewed data, so that we learn the network directly from the flow-cytometry data without any log or such transforms. Our pre-processing is limited to shifting the data for each protein so that it consists of non-negative values.

We then learned an exponential graphical model from this flow cytometry data using stability selection [Liu et al. 2010] to select the sparsity of the graph. The estimated protein-signaling network is shown on the right in Figure 2. Recall that the exponential graphical model restricts the edge weights to be non-negative; because of the negative inverse link, this entails that only negative associations can be estimated. Thus, our estimated network only finds inhibitory relationships among proteins. Our results indicate that PKA, protein kinase A, is a major protein inhibitor in cell signaling networks. This protein was also found to be a well-connected hub using both Bayesian networks [Sachs et al. 2005] and the graphical lasso [Friedman et al. 2007]. Additionally, our estimated neighborhood of PKA consists of five proteins which are consistent with those negative conditional dependencies also found by the graphical lasso algorithm.

5 Discussion

We have studied the class of graphical models that arise when we assume that node-wise conditional distributions follow an exponential family distribution. We have provided simple M-estimators for learning the network by fitting node-wise penalized GLMs that enjoy strong statistical recovery properties. Our work has broadened the class of off-the-shelf graphical models from the Ising model and Gaussian graphical model to encompass a wider range of parametric distributions. These classes of graphical models may be of further interest to the statistical community as they provide closed form multivariate densities as extensions of several univariate exponential family distributions (e.g. Poisson, exponential, negative binomial) where few cur-
rently exist. Furthermore, the statistical analysis of our M-estimator required subtle techniques that may be of general interest in the analysis of sparse M-estimation.

There are many avenues for future work in studying this model for specific distributional families. In particular, our model sometimes places restrictions on the parameter space. A key question is whether these restrictions could be relaxed for specific exponential family distributions. Additionally, we have focused on families with linear sufficient statistics (e.g. Gaussian, Bernoulli, Poisson, exponential, negative binomial); our models can be studied with non-linear sufficient statistics or multi-parameter distributions as well. Overall, our work has opened the door for learning Markov Networks from a broad class of distributions, the properties and applications of which leave much room for future research.

References

Abdelmohsen, K., Kim, M., Srikantan, S., Mercken, E., Brennan, S., Wilson, G., de Cabo, R., and Gorospe, M. (2010), “miR-519 suppresses tumor growth by reducing HuR levels,” Cell cycle (Georgetown, Tex.), 9, 1354.

Acemoglu, D. (2009), “The Crisis of 2008: Lessons for and from Economics,” Critical Review, 21, 185–194.

Allen, G. I. and Liu, Z. (2012), “A Log-Linear Graphical Model for Inferring Genetic Networks from High-Throughput Sequencing Data,” IEEE International Conference on Bioinformatics and Biomedicine.

Besag, J. (1974), “Spatial interaction and the statistical analysis of lattice systems,” Journal of the Royal Statistical Society. Series B (Methodological), 36, 192–236.

Bishop, Y., Fienberg, S., and Holland, P. (2007), Discrete multivariate analysis, Springer Verlag.

Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010), “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments,” BMC bioinformatics, 11, 94.

Cancer Genome Atlas Research Network (2012), “Comprehensive molecular portraits of human breast tumours,” Nature, 490, 61–70.

Clifford, P. (1990), “Markov random fields in statistics,” in Disorder in physical systems, eds. Grimmett, G. and Welsh, D. J. A., Oxford Science Publications.

Dobra, A. and Lenkoski, A. (2011), “Copula Gaussian graphical models and their application to modeling functional disability data,” The Annals of Applied Statistics, 5, 969–993.

Friedman, J., Hastie, T., and Tibshirani, R. (2007), “Sparse inverse covariance estimation with the lasso,” Biostatistics, 9, 432–441.
Friedman, N. (2004), “Inferring cellular networks using probabilistic graphical models,” *Science Signalling*.

Herzenberg, L., Tung, J., Moore, W., Herzenberg, L., and Parks, D. (2006), “Interpreting flow cytometry data: a guide for the perplexed,” *Nature immunology*, 7, 681–685.

Jalali, A., Ravikumar, P., Vasuki, V., and Sanghavi, S. (2011), “On Learning Discrete Graphical Models using Group-Sparse Regularization,” in *Inter. Conf. on AI and Statistics (AISTATS)*, 14.

Keklikoglou, I., Koerner, C., Schmidt, C., Zhang, J., Heckmann, D., Shavinskaya, A., Allgayer, H., Gückel, B., Fehm, T., Schneeweiss, A., et al. (2011), “MicroRNA-520/373 family functions as a tumor suppressor in estrogen receptor negative breast cancer by targeting NF-κB and TGF-β signaling pathways,” *Oncogene*.

Lafferty, J., Liu, H., and Wasserman, L. (2012), “Sparse Nonparametric Graphical Models,” *Arxiv preprint arXiv:1201.0794*.

Lauritzen, S. (1996), *Graphical models*, Oxford University Press, USA.

Leek, J., Scharpf, R., Bravo, H., Simcha, D., Langmead, B., Johnson, W., Geman, D., Baggerly, K., and Irizarry, R. (2010), “Tackling the widespread and critical impact of batch effects in high-throughput data,” *Nature Reviews Genetics*, 11, 733–739.

Li, J., Witten, D., Johnstone, I., and Tibshirani, R. (2011), “Normalization, testing, and false discovery rate estimation for RNA-sequencing data,” *Biostatistics*.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012a), “High Dimensional Semiparametric Gaussian Copula Graphical Models,” *Arxiv preprint arXiv:1202.2169*.

— (2012b), “The Nonparanormal skeptic,” *Arxiv preprint arXiv:1206.6488*.

Liu, H., Lafferty, J., and Wasserman, L. (2009), “The nonparanormal: Semiparametric estimation of high dimensional undirected graphs,” *The Journal of Machine Learning Research*, 10, 2295–2328.

Liu, H., Roeder, K., and Wasserman, L. (2010), “Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models,” *Arxiv preprint arXiv:1006.3316*.

Marioni, J., Mason, C., Mane, S., Stephens, M., and Gilad, Y. (2008), “RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays,” *Genome research*, 18, 1509–1517.

McCullagh, P. and Nelder, J. (1989), *Generalized linear models*, Monographs on statistics and applied probability 37, New York: Chapman and Hall/CRC.

Meinshausen, N. and Bühlmann, P. (2006), “High-dimensional graphs and variable selection with the Lasso,” *Annals of Statistics*, 34, 1436–1462.
Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. (2010), “A unified framework for high-dimensional analysis of \( M \)-estimators with decomposable regularizers.”

Pe’er, D., Regev, A., Elidan, G., and Friedman, N. (2001), “Inferring subnetworks from perturbed expression profiles,” *Bioinformatics (Oxford, England)*, 17, S215–S224.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2010), “High-dimensional Ising model selection using \( \ell_1 \)-regularized logistic regression,” *Annals of Statistics*, 38, 1287–1319.

Sachs, K., Perez, O., Pe’er, D., Lauffenburger, D., and Nolan, G. (2005), “Causal protein-signaling networks derived from multiparameter single-cell data,” *Science Signalling*, 308, 523.

Speed, T. P. and Kiiiveri, H. T. (1986), “Gaussian Markov distributions over finite graphs,” *Annals of Statistics*, 14, 138–150.

Varin, C., Reid, N., and Firth, D. (2011), “An Overview of Composite Likelihood Methods,” *Statistica Sinica*, 21, 5–42.

Varin, C. and Vidoni, P. (2005), “A note on composite likelihood inference and model selection,” *Biometrika*, 92, 519–528.

Wainwright, M. and Jordan, M. (2008), “Graphical models, exponential families, and variational inference,” *Foundations and Trends® in Machine Learning*, 1, 1–305.

Wainwright, M. J. (2009), “Sharp thresholds for high-dimensional and noisy sparsity recovery using \( \ell_1 \)-constrained quadratic programming (Lasso),” *IEEE Trans. Information Theory*, 55, 2183–2202.

Wei, Z. and Li, H. (2007), “A Markov random field model for network-based analysis of genomic data,” *Bioinformatics*, 23, 1537.

Xue, L. and Zou, H. (2012), “Regularized rank-based estimation of high-dimensional non-paranormal graphical models,” *Forthcoming in Ann. Statist*.

Yu, F., Yao, H., Zhu, P., Zhang, X., Pan, Q., Gong, C., Huang, Y., Hu, X., Su, F., Lieberman, J., et al. (2007), “let-7 regulates self renewal and tumorigenicity of breast cancer cells,” *Cell*, 131, 1109–1123.
Appendix

A Proof of Proposition 1

The proof follows the development in Besag (1974), where they consider the case with \( k = 2 \). Denote \( Q(x) \) as \( Q(X) = \log(P(X)/P(0)) \), for any \( X = (X_1, \ldots, X_p) \in \mathcal{X}^p \). Given any \( X \), also denote \( \bar{X}_s := (X_1, \ldots, X_{s-1}, 0, X_{s+1}, \ldots, X_p) \). Now, consider the following general form for \( Q(X) \):

\[
Q(X) = \sum_{t_1 \in V} X_{t_1} G_{t_1}(X_{t_1}) + \ldots + \sum_{t_1, \ldots, t_k \in V} X_{t_1} \ldots X_{t_k} G_{t_1, \ldots, t_k}(X_{t_1}, \ldots, X_{t_k}), \tag{24}
\]

since the joint distribution has at most factors of size \( k \). It can then be seen that

\[
\exp(Q(X) - Q(\bar{X}_1)) = P(X_s|X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_p)/P(0|X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_p), \tag{25}
\]

where the first equality follows from the definition of \( Q \), and the second equality follows from some algebra (See for instance Section 2 in Besag (1974)). Now, consider simplifications of both sides of (25). Given the form of \( Q(X) \) in (24), we have

\[
Q(X) - Q(\bar{X}_1) = \sum_{t=2}^{p} (X_t G_{tt}(X_t, X_t) + \sum_{t_2, \ldots, t_k \in \{2, \ldots, p\}} X_{t_2} \ldots X_{t_k} G_{t_2, \ldots, t_k}(X_{t_1}, \ldots, X_{t_k})) \tag{26}
\]

Also, given the exponential family form of the node-conditional distribution specified in the statement,

\[
\log \frac{P(X_s|X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_p)}{P(0|X_1, \ldots, X_{s-1}, X_{s+1}, \ldots, X_p)} = E(X_{V \setminus s})(B(X_s) - B(0)) + (C(X_s) - C(0)). \tag{27}
\]

Setting \( X_t = 0 \) for all \( t \neq s \) in (25), and using the expressions for the left and right hand sides in (26) and (27), we obtain,

\[
X_s G_s(X_s) = E(0)(B(X_s) - B(0)) + (C(X_s) - C(0)). \tag{28}
\]

Setting \( X_r = 0 \) for all \( r \notin \{s, t\} \),

\[
X_s G_s(X_s) + X_s X_t G_{st}(X_s, X_t) = E(0, \ldots, X_t, \ldots, 0)(B(X_s) - B(0)) + (C(X_s) - C(0)). \tag{29}
\]

Similarly,

\[
X_t G_t(X_t) + X_s X_t G_{st}(X_s, X_t) = E(0, \ldots, X_s, \ldots, 0)(B(X_t) - B(0)) + (C(X_t) - C(0)). \tag{30}
\]
From the above three equations, we obtain:

$$X_sX_tG_{st}(X_s, X_t) = \theta_{st}(B(X_s) - B(0))(B(X_t) - B(0)).$$

More generally, by considering non-zero triplets, and setting $X_r = 0$ for all $r \not\in \{s, t, u\}$, we obtain,

$$X_sG_s(X_s) + X_sX_tG_{st}(X_s, X_t) + X_sX_uG_{su}(X_s, X_u) + X_sX_tX_uG_{stu}(X_s, X_t, X_u) = E(0, \ldots, X_t, \ldots, X_u, \ldots, 0)(B(X_s) - B(0)) + (C(X_s) - C(0)),$$

so that by a similar reasoning we can obtain

$$X_sX_tX_uG_{stu}(X_s, X_t, X_u) = \theta_{stu}(B(X_s) - B(0))(B(X_t) - B(0))(B(X_u) - B(0)).$$

More generally, we can show that

$$X_{t_1} \ldots X_{t_k} G_{t_1, \ldots, t_k}(X_{t_1}, \ldots, X_{t_k}) = \theta_{t_1, \ldots, t_k}(B(X_{t_1}) - B(0)) \ldots (B(X_{t_k}) - B(0)).$$

Thus, the $k$-th order factors in the joint distribution as specified in (24) are tensor products of $(B(X_s) - B(0))$, thus proving the statement of the proposition.

### B Proof of Proposition [2]

**Proof.** By the definition of (20) with the following simple calculation, the moment generating function of $X^2_s$ becomes:

$$\log \mathbb{E}[\exp(tX^2_s)] = \log \int_{\mathcal{X}^p} \exp \left\{ tX^2_s + \sum_{s \in V} \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t + \sum_{s \in V} C(X_s) \right\} \nu(dx)$$

$$= \bar{A}_s(t; \theta^*) - \bar{A}_s(0; \theta^*).$$

Suppose that $t \leq 1$. Then, by a Taylor Series expansion, we have for some $r \in [0, 1]$,

$$\bar{A}_s(t; \theta^*) - \bar{A}_s(0; \theta^*) = \frac{\partial \bar{A}_s(\eta; \theta^*)}{\partial \eta}(0) + \frac{1}{2} t^2 \frac{\partial^2 \bar{A}_s(\eta; \theta^*)}{\partial \eta^2}(rt)$$

$$\leq \kappa_v t + \frac{1}{2}\kappa_h t^2,$$

where the inequality uses Condition [3]. Note that since the derivative of log-partition function is the mean of the corresponding sufficient statistics and $\bar{A}_s(0; \theta) = A(\theta)$, $\frac{\partial \bar{A}_s(\eta; \theta^*)}{\partial \eta}(0) = \mathbb{E}[X^2_s] \leq \kappa_v$ by assumption. Thus, again by the standard Chernoff bounding technique, for $t \leq 1$,

$$P \left( \frac{1}{n} \sum_{i=1}^n (X^{(i)}_s)^2 \geq \delta \right) \leq \exp(-n\delta t + n\kappa_v t + \frac{n}{2}\kappa_h t^2)$$

$$\leq \exp(-n (\delta - \kappa_v)^2 \frac{2\kappa_h}{2\kappa_h}) \leq \exp(-n \frac{\delta^2}{4\kappa_h^2}),$$

for $\delta \leq \min\{2\kappa_v/3, \kappa_h + \kappa_v\}$.  

\qed
C Proof of Proposition 3

Proof. Let $\bar{v} \in \mathbb{R}^{p+1}$ be the zero-padded parameter with only one non-zero coordinate, which is 1, for the sufficient statistics $X_s$ so that $\|\bar{v}\|_2 = 1$. A simple calculation shows that

$$\log \mathbb{E}[\exp(X_s)] = A(\theta^*) + \bar{v} - A(\theta^*).$$

By a Taylor Series expansion and Condition 3, we have for some $r$

$$V$$

where the inequality (i) uses the fact that $\bar{v}$ has only nonzero element for the sufficient statistics $X_s$. Thus, again by the standard Chernoff bounding technique, for any positive constant $a$, $P(X_s \geq a) \leq \exp(-a + \kappa_m + \frac{1}{2}\kappa_h)$, and by setting $a = \delta \log \eta$ we have

$$P(X_s \geq \delta \log \eta) \leq \exp(-\delta \log \eta + \kappa_m + \frac{1}{2}\kappa_h) \leq c\eta^{-\delta},$$

where $c = \exp(\kappa_m + \frac{1}{2}\kappa_h)$, as claimed. $\square$

D Proof of Lemma 2

Proof. For a fixed $t \in \{1, ..., p - 1\}$, we define $V_t^{(i)}$ for notational convenience so that

$$W^n_t = \frac{1}{n} \sum_{i=1}^{n} X_s^{(i)} X_t^{(i)} - X_t^{(i)} D'(\theta^*_s + \langle \theta^*_s, X_t^{(i)} \rangle) = \frac{1}{n} \sum_{i=1}^{n} V_t^{(i)}$$

Consider the upper bound on the moment generating function of $V_t^{(i)}$, conditioned on $X_{V\setminus s}$,

$$\mathbb{E}[\exp(t'V_t^{(i)}|X_{V\setminus s})] = \sum_{X_t^{(i)}} \exp \left\{ t' \left[ X_s^{(i)} X_t^{(i)} - X_t^{(i)} D'(\theta^*_s + \langle \theta^*_s, X_t^{(i)} \rangle) \right] \right\}$$

$$+ (C(X_s^{(i)} + X_t^{(i)} \theta^*_s + \langle \theta^*_s, X_t^{(i)} \rangle) - D(\theta^*_s + \langle \theta^*_s, X_t^{(i)} \rangle))$$

$$= \exp \left\{ t' \left[ \theta^*_s + \langle \theta^*_s, X_t^{(i)} \rangle + t' X_t^{(i)} - D(\theta^*_s + \langle \theta^*_s, X_t^{(i)} \rangle) - t' X_t^{(i)} D'(\theta^*_s + \langle \theta^*_s, X_t^{(i)} \rangle) \right] \right\}$$

$$= \exp \left\{ \frac{t^2}{2} X_t^{(i)} D''(\theta^*_s + \langle \theta^*_s, X_t^{(i)} \rangle + v_i t' X_t^{(i)}) \right\}$$

for some $v_i \in [0, 1]$. 27
where the last equality holds by the second-order Taylor series expansion. Consequently, we have
\[
\frac{1}{n} \sum_{i=1}^{n} \log E[\exp(t'V_t^{(i)})|X_{V_{\backslash s}}^{(i)}] \leq \frac{1}{n} \sum_{i=1}^{n} \frac{t'^2}{2} \left(X_t^{(i)}\right)^2 D''(\theta^*_s + \langle \theta^*_s, X_{V_{\backslash s}}^{(i)} \rangle + v_i t' X_t^{(i)}).
\]

First, we define the event: \(\xi_1 := \left\{ \text{max}_{i,s} |X_s^{(i)}| \leq 4 \log p' \right\}\). Then, by Proposition 3 we obtain \(P[\xi_1^c] \leq c_1 n p^{t' - 4} \leq c_1 p'^{-4}\). Provided that \(t' \leq \kappa_2(n, p)\), we can use Condition 4 to control the second-order derivative of log-partition function and we obtain
\[
\frac{1}{n} \sum_{i=1}^{n} \log E[\exp(t'V_t^{(i)})|X_{V_{\backslash s}}^{(i)}] \leq \frac{\min \{2\kappa_v/3, 2\kappa_h + \kappa_v\} \log n}{4}\] with probability at least \(1 - p'^{-2}\). Now, for each index \(t\), the variables \(\frac{1}{n} \left\{ \left(X_t^{(i)}\right)^2 \right\}_{i=1}^{n}\) satisfy the tail bound in Proposition 2. Let us define the event \(\xi_2 := \left\{ \text{max}_{t=1,...,p-1} \frac{1}{n} \sum_{i=1}^{n} \left(X_t^{(i)}\right)^2 \right\} \leq \kappa_4\) for some constant \(\kappa_4 \leq \min \{2\kappa_v/3, 2\kappa_h + \kappa_v\}\). Then, we can establish the upper bound of probability \(P[\xi_2^c]\) by a union bound,
\[
P[\xi_2^c] \leq \exp\left(-32 n \log p\right) \leq \exp(-c_2 n)
\]
as long as \(n \geq \frac{\kappa_4}{\kappa_4^2} \log p\). Therefore, conditioned on \(\xi_1, \xi_2\), the moment generating function is bounded as follows:
\[
\frac{1}{n} \sum_{i=1}^{n} \log E[\exp(t'V_t^{(i)})|X_{V_{\backslash s}}^{(i)}, \xi_1, \xi_2] \leq \frac{\kappa_1(n, p) \kappa_4 t'^2}{2} \quad \text{for } t' \leq \kappa_2(n, p).
\]
The standard Chernoff bound technique implies that for any \(\delta > 0\),
\[
P\left[ \frac{1}{n} \sum_{i=1}^{n} |V_t^{(i)}| > \delta | \xi_1, \xi_2 \right] \leq 2 \exp \left( n \frac{\kappa_1(n, p) \kappa_4 t'^2}{2} - t' \delta \right) \quad \text{for } t' \leq \kappa_2(n, p).
\]
Setting \(t' = \frac{\delta}{\kappa_1(n, p) \kappa_4}\):
\[
P\left[ \frac{1}{n} \sum_{i=1}^{n} |V_t^{(i)}| > \delta | \xi_1, \xi_2 \right] \leq 2 \exp \left( -\frac{n \delta^2}{2 \kappa_1(n, p) \kappa_4} \right) \quad \text{for } \delta \leq \kappa_1(n, p) \kappa_2(n, p) \kappa_4.
\]
Suppose that \(\frac{\alpha}{2 - \alpha} \frac{\lambda_n}{4} \leq \kappa_1(n, p) \kappa_2(n, p) \kappa_4\) for large enough \(n\); thus setting \(\delta = \frac{\alpha}{2 - \alpha} \frac{\lambda_n}{4}\):
\[
P\left[ \frac{1}{n} \sum_{i=1}^{n} |V_t^{(i)}| > \frac{\alpha}{2 - \alpha} \frac{\lambda_n}{4} | \xi_1, \xi_2 \right] \leq 2 \exp \left( -\frac{\alpha^2}{(2 - \alpha)^2} \frac{n \lambda_n^2}{32 \kappa_1(n, p) \kappa_4} \right),
\]
28
and by a union bound, we obtain
\[
P\left[\|W^n\|_\infty > \frac{\alpha}{2 - \alpha} \frac{\lambda_n}{4} \mid \xi_1, \xi_2\right] \leq 2 \exp\left(-\frac{\alpha^2}{(2 - \alpha)^2} \frac{n\lambda_n^2}{32\kappa_1(n,p)\kappa_4} + \log p\right).
\]

Finally, provided that \(\lambda_n \geq \frac{8(2 - \alpha)}{\alpha} \sqrt{\kappa_1(n,p)\kappa_4} \sqrt{\frac{\log p}{n}}\), we obtain
\[
P\left[\|W^n\|_\infty > \frac{\alpha}{2 - \alpha} \frac{\lambda_n}{4}\right] \leq c_1p^{r-2} + \exp(-c_2n) + \exp(-c_3n),
\]
where we use the fact that \(P(X) \leq P(\xi_1^c) + P(\xi_2^c) + P(X|\xi_1, \xi_2)\).

\[\square\]

**E Proof of Lemma 3**

*Proof*. In order to establish the error bound \(|\hat{\theta}_S - \theta^*_S|_2 < B\) for some radius \(B\), several works (e.g. [Negahban et al. (2010); Ravikumar et al. (2010)]) proved that it suffices to show \(F(u_S) > 0\) for all \(u_S := \hat{\theta}_S - \theta^*_S\) s.t. \(\|u_S\|_2 = B\) where
\[
F(u_S) := \ell(\theta^*_S + u_S; X^n_i) - \ell(\theta^*_S; X^n_i) + \lambda_n(\|\theta^*_S + u_S\|_2 - \|\theta^*_S\|_2).
\]

Note that for \(\hat{u}_S := \hat{\theta}_S - \theta^*_S\), \(F(\hat{u}_S) = 0\). From now on, we show that \(F(u_s)\) is strictly positive on the boundary of the ball with radius \(B = M\lambda_n\sqrt{d}\) where \(M > 0\) is a parameter that we will choose later in this proof. Some algebra yields
\[
F(u_s) \geq (\lambda_n\sqrt{d})^2 \left\{ -\frac{1}{4} M + q^*M^2 - M \right\}
\]
where \(q^*\) is the minimum eigenvalue of \(\nabla^2\ell(\theta^*_S + vu_S; X^n_i)\) for some \(v \in [0, 1]\). Moreover,
\[
q^* := \lambda_{\min}(\nabla^2\ell(\theta^*_S + vu_S)) \\
\geq \min_{v \in [0,1]} \lambda_{\min}(\nabla^2\ell(\theta^*_S + vu_S)) \\
\geq \lambda_{\min}\left[\frac{1}{n} \sum_{i=1}^n D''(\theta^*_S + \langle \theta^*_S, X^{(i)}_S \rangle)X^{(i)}_S(X^{(i)}_S)^T \right] - \max_{v \in [0,1]} \left\|\frac{1}{n} \sum_{i=1}^n D''((\theta^*_S + vu_s) + \langle \theta^*_S + vu_s, X^{(i)}_S \rangle)(u^T_S X^{(i)}_S)X^{(i)}_S(X^{(i)}_S)^T \right\|_2 \\
\geq \lambda_{\min} - \max_{v \in [0,1]} \sum_{y \in [0,1]} \left\|D''((\theta^*_S + vu_s) + \langle \theta^*_S + vu_s, X^{(i)}_S \rangle) \right\| \left\|\langle u_s, X^{(i)}_S \rangle \right\| \left(\langle X^{(i)}_S, y \rangle \right)^2
\]
where \(y \in \mathbb{R}^d\) s.t \(\|y\|_2 = 1\). Similarly as in the previous proof, we consider the event \(\xi_1\) with probability at least \(1 - c_1p^{r-2}\). Then, since all the elements in vector \(X^{(i)}_S\) is
smaller than $4 \log p' \cdot |\langle u_S, X_S^{(i)} \rangle| \leq 4 \log p' \sqrt{d} \|u_S\|_2 = 4 \log p' M \lambda_n d$ for all $i$. At the same time, by Condition 4 \[ |D''((\theta^*_S + vu_S) + (\theta^*_S + vu_S, X_S^{(i)}))| \leq \kappa_3(n, p). \] Note that $\theta^*_S + vu_S$ is a convex combination of $\theta^*_S$ and $\bar{\theta}_S$, and as a result, we can directly apply the Condition 4. Hence, conditioned on $\xi_1$, we have

$$q^* \geq \lambda_{\text{min}} - 4 \lambda_{\text{max}} M \lambda_n d \kappa_3(n, p) \log p'.$$

As a result, assuming that $\lambda_n \leq \frac{\lambda_{\text{min}}}{8 \lambda_{\text{max}} M \kappa_3(n, p) \log p'}$, $q^* \geq \frac{\lambda_{\text{min}}}{2}$. Finally, from (32), we obtain

$$F(u_s) \geq (\lambda_n \sqrt{d})^2 \left\{ - \frac{1}{4} M + \frac{\lambda_{\text{min}}}{2} M_2 - M \right\},$$

which is strictly positive for $M = \frac{5}{\lambda_{\text{min}}}$. Therefore, if $\lambda_n \leq \frac{\lambda_{\text{min}}}{8 \lambda_{\text{max}} M \kappa_3(n, p) \log p'} \leq \frac{\lambda_{\text{min}}^2}{40 \lambda_{\text{max}} \kappa_3(n, p) \log p'}$, then

$$\|\hat{\theta}_S - \theta^*_S\|_2 \leq \frac{5}{\lambda_{\text{min}}} \sqrt{d} \lambda_n,$$

which completes the proof. \(\square\)

## F Proof of Lemma 4

**Proof.** In the proof, we are going to show that $\|R^n\|_{\infty} \leq 4 \kappa_3(n, p) \log p' \lambda_{\text{max}} \|\hat{\theta}_S - \theta^*_S\|_2$. Then, since the conditions of Lemma 4 are stronger than those of Lemma 3, from the result of Lemma 3 we can conclude that

$$\|R^n\|_{\infty} \leq \frac{100 \kappa_3(n, p) \lambda_{\text{max}} \log p' \lambda_n^2 d}{\lambda_{\text{min}}^2},$$

as claimed in Lemma 4.

From the definition of $R^n$, for a fixed $t \in \{1, \ldots, p-1\}$, $R^n_t$ can be written as

$$\frac{1}{n} \sum_{i=1}^n \left[ D''(\bar{\theta}_{\text{\hat{\theta}}}, \{\bar{\theta}_{\text{\hat{\theta}}}, X^{(i)}_{V \setminus \hat{\theta}}\}) - D''(\bar{\theta}_{\text{\hat{\theta}}}, \{\tilde{\theta}_{\text{\hat{\theta}}}, X^{(i)}_{V \setminus \hat{\theta}}\}) \right] [X^{(i)}_{V \setminus \hat{\theta}}(X^{(i)}_{V \setminus \hat{\theta}})^T]_t [\hat{\theta} - \theta^*]$$

where $\tilde{\theta}^{(i)}_{\text{\hat{\theta}}}$ is some point in the line between $\bar{\theta}_{\text{\hat{\theta}}}$ and $\theta^*_s$, i.e., $\tilde{\theta}^{(i)}_{\text{\hat{\theta}}} = v_t \bar{\theta}_{\text{\hat{\theta}}} + (1 - v_t) \theta^*_s$ for $v_t \in [0, 1]$. By another application of the mean value theorem, we have

$$R^n_t = -\frac{1}{n} \sum_{i=1}^n \left\{ D''(\bar{\theta}_{\text{\hat{\theta}}}, \{\bar{\theta}_{\text{\hat{\theta}}}, X^{(i)}_{V \setminus \hat{\theta}}\}) \right\} \left\{ v_t [\bar{\theta}_{\text{\hat{\theta}}} - \theta^*_s]^T X^{(i)}_{V \setminus \hat{\theta}} (X^{(i)}_{V \setminus \hat{\theta}})^T [\bar{\theta}_{\text{\hat{\theta}}} - \theta^*_s] \right\}$$

for a some point $\bar{\theta}^{(i)}_{\text{\hat{\theta}}}$ between $\tilde{\theta}^{(i)}_{\text{\hat{\theta}}}$ and $\theta^*_s$. Similarly in the previous proofs, conditioned on the event $\xi_1$, we obtain

$$|R^n| \leq \frac{4 \kappa_3(n, p) \log p'}{n} \sum_{i=1}^n \left\{ v_t [\bar{\theta}_{\text{\hat{\theta}}} - \theta^*_s]^T X^{(i)}_{V \setminus \hat{\theta}} (X^{(i)}_{V \setminus \hat{\theta}})^T [\bar{\theta}_{\text{\hat{\theta}}} - \theta^*_s] \right\}.$$
Performing some algebra yields
\[ |R_t^n| \leq 4\kappa_3(n, p)\lambda_{\max} \log p' \|\hat{\theta}_S - \theta^*_S\|_2^2, \] for all \( t \in \{1, ..., p - 1\} \)
with probability at least \( 1 - c_1p'^{-2} \), which completes the proof. \( \square \)