Electronic Journal of Statistics
Vol. 8 (2014) 2242–2263
ISSN: 1935-7524
DOI: 10.1214/14-EJS946

Estimation of high-dimensional partially-observed discrete Markov random fields

Yves F. Atchade[*]

University of Michigan, 1085 South University
Ann Arbor, 48109, MI, United States
e-mail: yvesa@umich.edu

Abstract: We consider the problem of estimating the parameters of discrete Markov random fields from partially observed data in a high-dimensional setting. Using a ℓ^1-penalized pseudo-likelihood approach, we fit a misspecified model obtained by ignoring the missing data problem. We derive an estimation error bound that highlights the effect of the misspecification. We report some simulation results that illustrate the theoretical findings.

MSC 2010 subject classifications: 62M40, 62G20.
Keywords and phrases: Network estimation, high-dimensional inference, penalized likelihood inference, misspecification, pseudo-likelihood, Markov random fields.

Received August 2013.

Contents

[*] This work is partly supported by NSF grant DMS 0906631.

1. Introduction and statement of the results

The problem of estimating high-dimensional networks has recently attracted a lot of attention in statistics and machine learning. Both in the continuous case using Gaussian graphical models [8, 13, 19, 7, 4, 17, 12], and in the discrete case using Markov random fields [2, 11, 16, 10]. This paper focuses mainly on discrete Markov random fields (MRF). Let $(X^{(1)}, \ldots, X^{(n)})$ be n i.i.d. random variables where $X^{(i)} = (X_1^{(i)}, \ldots, X_p^{(i)})$ is a p-dimensional vector of dependent random variables with joint density

$$f_\theta(x_1, \ldots, x_p) = \frac{1}{Z_\theta} \exp\left\{ \sum_{1 \leq s < s' \leq p} \theta(s, s') B(x_s, x_{s'}) \right\}, \tag{1.1}$$

for some symmetric function $B : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$, where X is a finite set. The real-valued symmetric matrix $\theta = \{\theta(s, s'),\ 1 \leq s < s' \leq p\}$ is the network structure and is the parameter of interest. The term Z_θ is a normalizing constant. This type of statistical models was pioneered by J. Besag [3] under the name auto-model. The nice feature of model (1.1) is that for any $1 \leq s \leq p$, the conditional density of X_s given $\{X_j, j \neq s\} = x \in \mathsf{X}^{p-1}$ is

$$f_\theta^{(s)}(u|x) = \frac{1}{Z_\theta^{(s)}} \exp\left\{ \sum_{j \neq i} \theta(s, j) B(u, x_j) \right\}, \tag{1.2}$$

for a normalizing constant $Z_\theta^{(s)} = Z_\theta^{(s)}(x)$. Therefore, $\theta(s, j) = 0$ implies that X_s and X_j are conditionally independent given the other variables X_k, $k \notin \{s, j\}$. Thus estimating θ provides us with the dependence structure and the magnitude of the dependence between these variables.

This paper focuses on the situation where the outcomes $X_j^{(i)}$ take discrete values (X is a finite set), although extension to a more general setting is possible without much difficulty. A number of recent work have shown that based on $(X^{(1)}, \ldots, X^{(n)})$, the true network structure denoted $\theta_\star$ can be consistently estimated using a number of methods, even when the number of entries of $\theta_\star$ is much large than n [11, 16, 10]. For computational tractability, a pseudo-likelihood approach is often preferred, even though this approach incurs a certain lost of efficiency. Working mainly with the auto-logistic model (where $\mathsf{X} = \{0, 1\}$, $B(u, v) = uv$), [10] shows that the ℓ^2-norm estimation error of the penalized pseudo-likelihood estimator is bounded from above by $\tau^{-1}\sqrt{a \log d/n}$, where a is the number of non-zero elements of $\theta_\star$ and τ is the smallest eigenvalue of the information matrix. [16] obtained similar results for a one-node-at-the-time ℓ^1-penalized pseudo-likelihood estimator. [18] also derived some properties of the oracle estimator with the SCAD penalty.

In many situations where network estimation is needed, the network data can be only partially observed because certain nodes are missing from the sample. For example, in social network analysis, some close friends or siblings might not be part of the survey. As another example, in protein-protein networks, the analysis is often restricted to the specific subgroup of proteins that is believed to carry a role in a given biological function. So doing, some important proteins might be omitted from the analysis. In the Gaussian case the distribution of the observed nodes remains Gaussian, but its conditional independence structure can be substantially altered by the missing data problem. [6] considered this issue and studied the problem of recovering the conditional independence structure among the observed nodes (as defined in the complete data setting). They address the issue by approximating the inverse covariance matrix of the observed nodes by a sum of a sparse matrix and a low-rank matrix. Key to their approach is the fact that the marginal distribution of the observed nodes remains Gaussian, albeit one with an altered covariance matrix. Under some regularity and identifiability conditions, these authors show that the sparse component of their model consistently estimates the covariance matrix (as defined in the complete data setting) between the observed nodes.

This paper consider the same issue for discrete MRF. Unlike the Gaussian case, discrete Markov random field distributions are not closed under marginalization. For example, if there exist r additional nodes denoted $p{+}1,\ldots,p{+}r$ such that the joint distribution of $(X_1,\ldots,X_p,X_{p+1},\ldots,X_{p+r})$ is an auto-model with network structure $\{\theta(s,s'),\ 1\leq s<s'\leq p+r\}$, then the joint (marginal) distribution of $(X_1,\ldots,X_p)$ is <u>not</u> of the form (1.1) in general. To take a specific example, if $r=1$ and $B(x,y)=B(x)B(y)$, then the joint (marginal) distribution of $(X_1,\ldots,X_p)$ is the mixture distribution

$$f_\theta(x_1,\ldots,x_p)=\frac{1}{C_\theta}\sum_{i\in\mathsf{X}}\exp\left\{\sum_{s=1}^{p}\theta_i(s)B(x_s)+\sum_{1\leq s<s'\leq p}\theta(s,s')B(x_s)B(x_{s'})\right\},$$

where $\theta_i(s)=B(i)\theta(s,p+1)$. The conditional distributions are also altered. Indeed, and keeping with the assumption $r=1$, if $|\theta(s,p+1)|>0$, then the conditional density of X_s given $\{X_\ell,\ \ell\neq s,\ 1\leq\ell\leq p\}$ depends not only on X_ℓ for all ℓ such that $|\theta(s,\ell)|>0$, but also on X_k for all k such that $|\theta(k,p{+}1)|>0$. Because the marginal distribution of the observed node belongs to a different family than (1.1), it seems unlikely that the “sparse + low-rank” approach of [6] would be of much use in this context. We propose to ignore the missing nodes and fit (the misspecified) model (1.1) to the observed data. It seems plausible that the resulting estimator would still be well-behaved to the extent that the missing data problem is limited. The goal of the paper is to formalize this idea.

We consider a large (possibly infinite) Markov random field model, where only part of the field is observed, and fit the misspecified model (1.1) using penalized pseudo-likelihood approach. We study conditions under which this procedure can recover the true network parameter. We show that the ℓ^2-norm estimation error of the procedure is at most $\tau^{-1}\sqrt{a}(\sqrt{\log d/n}+b)$, up to a multiplicative constant factor, where d (resp. a) is the number of possible edges (resp. the number of non-zeros entries) of the true network, τ is the smallest eigenvalue of the information matrix, and where the term b represents the effect of the missing nodes (see Theorem 1.2 for the exact statement). We conclude that the estimator $\hat{\theta}_n$ is robust to a small to moderate amount of missing data. We report some simulation results that are consistent with these findings. In practical situations where MRF are used, it is often unclear whether one is dealing with a partially observed field with important missing nodes. The above discussion thus stresses the need for methods of detecting the existence of missing nodes in Markov random field data. We leave this problem for future research.

The remainder of the paper is organized as follows. We define the model and estimator in Section 1.1, followed by the statement of the main result in Section 1.2. A simulation example is presented in Section 1.3. Section 2 develops the technical proofs.

1.1. The setting

Let X be a finite set. X is the sample space of the observations. Let $\mathcal{S}$ be a finite non-empty or countably infinite set that we assume, without any loss of generality to be a subset of the integer set $\mathbb{N}$. The set $\mathcal{S}$ represents the nodes of the network. We use the notation $\underline{\mathcal{S}}^2 \overset{\text{def}}{=} \{(s,\ell) \in \mathcal{S} \times \mathcal{S} :\; s < \ell\}$, the set of all ordered pairs of $\mathcal{S}$. More generally, if Λ is a subset of $\mathcal{S}$, we denote by $\underline{\Lambda}^2$, the set of all ordered pairs $(u,v) \in \Lambda \times \Lambda$, with $u < v$.

Let $B :\; \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ be a measurable function such that $B(x,y) = B(y,x)$ (symmetry). Throughout the paper we define

$$c_1 \overset{\text{def}}{=} \sup_{x,y,z \in \mathsf{X}} |B(y,x) - B(z,x)| < \infty, \tag{1.3}$$

which plays the role of the variance of the interaction statistics $B(X_s, x)$.

For a matrix $\theta :\; \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ and $s \in \mathcal{S}$, the θ-neighborhood of s is the set

$$\partial_\theta s \overset{\text{def}}{=} \{\ell \in \mathcal{S} :\; \ell \neq s \text{ and } |\theta(s,\ell)| > 0\},$$

and the θ-degree of node s is the (possibly infinite) quantity

$$\mathsf{deg}_\theta(s) \overset{\text{def}}{=} \sum_{\ell \in \mathcal{S} \setminus \{s\}} |\theta(s,\ell)| = \sum_{\ell \in \partial_\theta s} |\theta(s,\ell)|.$$

We denote $\mathcal{M}(\mathcal{S})$ the space of all symmetric matrices $\theta :\; \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ such that $\mathsf{deg}_\theta(s) < \infty$ for all $s \in \mathcal{S}$. For $\theta \in \mathcal{M}(\mathcal{S})$, let μ_θ be the probability measure on $(\mathsf{X}^{\mathcal{S}}, \mathcal{E}^{\mathcal{S}})$ such that if $X = \{X_s,\; s \in \mathcal{S}\}$ has distribution μ_θ, then the conditional distribution of X_s given $\{X_\ell,\; \ell \neq s\}$ exists and has probability mass function $f_\theta^{(s)}(\cdot|x)$, where for $u \in \mathsf{X}$, $x \in \mathsf{X}^{\mathcal{S}\setminus\{s\}}$,

$$f_\theta^{(s)}(u|x) = \frac{1}{Z_\theta^{(s)}} \exp\left\{ \sum_{\ell \in \mathcal{S}\setminus\{s\}} \theta(s,\ell) B(u, x_\ell) \right\}, \tag{1.4}$$

for a normalizing constant $Z_\theta^{(s)} = Z_\theta^{(s)}(x)$. Notice that $f_\theta^{(s)}(u|x)$ actually depends only on $x_{\partial_\theta s} \overset{\text{def}}{=} \{x_\ell :\ \ell \in \partial_\theta s\}$. As a result, we will interchangly write $f_\theta^{(s)}(u|x)$ and $f_\theta^{(s)}(u|x_{\partial_\theta s})$ to mean the same object. We call the process $\{X_s,\ s \in \mathcal{S}\}$ an auto-model Markov random field. We will take for granted that such distributions μ_θ exist. Obviously, this is the case if $\mathcal{S}$ is finite. In the case where $\mathcal{S}$ is infinite, it can be shown that μ_θ exists for any $\theta \in \mathcal{M}(\mathcal{S})$. This follows for instance from [9], Theorem 4.23 (a).

For $\theta_\star \in \mathcal{M}(\mathcal{S})$, let $\{X^{(i)},\ 1 \leq i \leq n\}$ be a sequence of i.i.d. auto-model Markov random fields with distribution $\mu_{\theta_\star}$ defined on some probability space with probability measure $\check{\mathbb{P}}_\star$ and expectation operator $\check{\mathbb{E}}_\star$. Let $\mathcal{D}$ be a finite subset of $\mathcal{S}$ with cardinality p. We assume that the random fields $X^{(i)}$ are only observed over $\mathcal{D}$, giving rise to observations $(X^{(1)}, \ldots, X^{(n)})$, where $X^{(i)} = \{X_k^{(i)},\ k \in \mathcal{D}\}$. We are interested in inferring the network parameter $\{\theta_\star(s,\ell),\ (s,\ell) \in \underline{\mathcal{D}}^2\}$ from the observed data.

Let $d \overset{\text{def}}{=} p(p-1)/2$ and denote $\mathcal{M}(\mathcal{D})$ the set of all symmetric finite matrices $\{\theta(s,\ell),\ (s,\ell) \in \underline{\mathcal{D}}^2\}$, that we identify with $\mathbb{R}^d$. For $s \in \mathcal{S}$, we define $\partial_s \overset{\text{def}}{=} \partial_{\theta_\star} s$ and called it the (true) neighborhood of s. We also define $\mathcal{D}_s \overset{\text{def}}{=} \mathcal{D} \setminus \{s\}$. Since the neighborhood system $\{\partial_s,\ s \in \mathcal{S}\}$ is not known, we introduce the approximate conditional distributions

$$\bar{f}_\theta^{(s)}(u|x) = \bar{f}_\theta^{(s)}(u|x_{\mathcal{D}_s}) \overset{\text{def}}{=} \frac{1}{\bar{Z}_\theta^{(s)}} \exp\left(\sum_{\ell \in \mathcal{D}_s} \theta(s,\ell) B(u, x_\ell) \right),\quad u \in \mathsf{X},\ x \in \mathsf{X}^{\mathcal{S}\setminus\{s\}}, \tag{1.5}$$

for some normalizing constant $\bar{Z}_\theta^{(s)}$. The difference between (1.5) and (1.4) is that $\bar{f}_\theta^{(s)}(u|x)$ depends only on the nodes in $\mathcal{D}$. In particular, $\bar{f}_\theta^{(s)}(u|x)$ depends on θ only through $\theta_s \overset{\text{def}}{=} \{\theta(s,\ell),\ \ell \in \mathcal{D}_s\}$. We define the functions

$$\ell_n(\theta) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \sum_{s \in \mathcal{D}} \log \bar{f}_\theta^{(s)}(X_s^{(i)} | X_{\mathcal{D}_s}^{(i)}), \quad \text{and}$$

$$Q_n(\theta) = \ell_n(\theta) - \lambda_n \sum_{(s,\ell) \in \underline{\mathcal{D}}^2} |\theta(s,\ell)|,\quad \theta \in \mathcal{M}(\mathcal{D}),$$

for some parameter $\lambda_n > 0$. Finally, we define

$$\mathsf{Argmax}\, Q_n \overset{\text{def}}{=} \{\theta \in \mathcal{M}(\mathcal{D}) :\ Q_n(\theta) = \sup_{\vartheta \in \mathcal{M}(\mathcal{D})} Q_n(\vartheta)\},$$

and we call any element $\hat{\theta}_n$ of $\mathsf{Argmax}\, Q_n$ a penalized pseudo-likelihood estimator of $\theta_\star$. It is useful to have some simple conditions under which $\mathsf{Argmax}\, Q_n$ well-defined. It is easy to see that the function Q_n is strictly concave. Thus if $\hat{\theta}_n$ exists, it is necessarily unique. The following result gives an easily verifiable condition under which $\hat{\theta}_n$ exists.

Proposition 1.1. *Suppose that for each $s \in \mathcal{D}$, there exists a finite constant $c(s)$ such that for all $\theta \in \mathcal{M}(\mathcal{D})$, all $u \in \mathsf{X}$ and for all $x \in \mathsf{X}^{\mathcal{D}_s}$,*

$$f_\theta^{(s)}(u|x) \le c(s).$$

Then $\hat{\theta}_n$ exists and is unique.

1.2. Non-asymptotic estimation error bound

In this section X denotes a Markov random field with distribution $\mu_{\theta_\star}$. Let $s \in \mathcal{D}$. Notice that if the entire $\theta_\star$-neighborhood of s (that is ∂s) is included in $\mathcal{D}$, then the approximate conditional distribution (1.5) and the true conditional distribution (1.4) would be the same: $\bar{f}_{\theta_\star}^{(s)}(\cdot|x_{\mathcal{D}_s}) = f_{\theta_\star}^{(s)}(\cdot|x_{\partial_s})$ for all $x \in \mathsf{X}^{\mathcal{S}\setminus\{s\}}$. In particular, we would have:

$$\mathbb{E}_\star\left[B(X_s, X_\ell)\right] = \mathbb{E}_\star\left[\int B(u, X_\ell)\bar{f}_{\theta_\star}^{(s)}(u|X_{\mathcal{D}_s})du\right],\ \ell \in \mathcal{D}_s.$$

This motivates the definition

$$b \overset{\text{def}}{=} \sup_{s\in\mathcal{D}}\sup_{\ell\in\mathcal{D}_s}\left|\mathbb{E}_\star\left[B(X_s, X_\ell)\right] - \mathbb{E}_\star\left[\int B(u, X_\ell)\bar{f}_{\theta_\star}^{(s)}(u|X_{\mathcal{D}_s})du\right]\right|. \tag{1.6}$$

The quantity b measures the effect of the missing nodes. It measures how well the misspecified conditional densities $\bar{f}_{\theta_\star}^{(s)}(u|x_{\mathcal{D}_s})$ approximate the correct conditional densities $f_{\theta_\star}^{(s)}(u|x_{\mathcal{S}\setminus\{s\}})$ in terms of matching the first moment of the statistics $B(X_s, x_\ell)$. As we will see below, the quantity b is the main effect of the misspecification on the recovery rate of the ℓ^1-penalized pseudo-likelihood estimator.

Let $I \overset{\text{def}}{=} \{(s,\ell) \in \underline{\mathcal{D}}^2 :\ \theta_\star(s,\ell) \neq 0\}$, and denote a the cardinality of I. Set

$$\Delta \overset{\text{def}}{=} \left\{\theta \in \mathcal{M}(\mathcal{D}) : \sum_{(s,\ell)\in\underline{\mathcal{D}}^2\setminus I}|\theta(s,\ell)| \le 3\sum_{(s,\ell)\in I}|\theta(s,\ell)|\right\}.$$

For $\theta \in \mathcal{M}(\mathcal{D})$, we introduce the semi-norm

$$\|\theta\|_{2\star} \overset{\text{def}}{=} \left\{\sum_{(s,\ell)\in I}|\theta(s,\ell)|^2\right\}^{1/2}.$$

For $s \in \mathcal{D}$, $\ell, \ell' \in \mathcal{D}_s$, we define the random variable

$$\begin{aligned} H^{(s)}(\ell,\ell';X) &\overset{\text{def}}{=} \int_{\mathsf{X}} B(u,X_\ell)B(u,X_{\ell'})\bar{f}_{\theta_\star}^{(s)}(u|X_{\mathcal{D}_s})du \\ &\quad - \left\{\int_{\mathsf{X}} B(u,X_\ell)\bar{f}_{\theta_\star}^{(s)}(u|X_{\mathcal{D}_s})du\right\}\left\{\int_{\mathsf{X}} B(u,X_{\ell'})\bar{f}_{\theta_\star}^{(s)}(u|X_{\mathcal{D}_s})du\right\} \end{aligned} \tag{1.7}$$

and we set

$$C^{(s)}(\ell,\ell') \stackrel{\text{def}}{=} \mathbb{E}_\star\left[H^{(s)}(\ell,\ell';X)\right].$$

The family of matrices $\{C^{(s)},\ s\in\mathcal{D}\}$ plays the role of information matrix. Clearly, each matrix $C^{(s)}$ defines a quadratic form on $\mathbb{R}^{p-1}$ by

$$\theta_s' C^{(s)}\theta_s \stackrel{\text{def}}{=} \sum_{\ell\in\mathcal{D}_s}\sum_{\ell'\in\mathcal{D}_s}\theta(s,\ell)\theta(s,\ell')C^{(s)}(\ell,\ell'),$$

where we write $\theta_s \stackrel{\text{def}}{=} \{\theta(s,\ell),\ \ell\in\mathcal{D}_s\}$. We impose the following restricted strong convexity-type assumption.

A1 There exists $\tau>0$ such that

$$\sum_{s\in\mathcal{D}}\theta_s' C^{(s)}\theta_s \geq 2\tau\|\theta\|_{2\star}^2,\quad \theta\in\Delta. \tag{1.8}$$

Theorem 1.2. *Assume A1 and take $\lambda_n \geq 4b+8c_1\sqrt{\frac{\log d}{n}}$. Suppose that $n\tau^2 \geq 2(64^2)c_1^2a^2\log(2d)$, and $48c_1a\lambda_n<\tau$. Then*

$$\left\|\hat{\theta}_n-\bar{\theta}_\star\right\|_2 \leq 26\sqrt{a}\lambda_n,$$

with a probability at least $1-\frac{3}{d}$, where $\bar{\theta}_\star = \{\theta_\star(s,\ell),\ (s,\ell)\in\underline{\mathcal{D}}^2\}$, and for $u\in\mathcal{M}(\mathcal{D})$, $\|u\|_2 \stackrel{\text{def}}{=} \{\sum_{(s,\ell)\in\underline{\mathcal{D}}^2}|u(s,\ell)|^2\}^{1/2}$.

Remark 1. Taking $\lambda_n = 4b+8c_1\sqrt{\frac{\log d}{n}}$, and assuming that $48c_1a\lambda_n<\tau$, the bound suggests that the convergence rate of the estimator $\hat{\theta}_n$ is $\tau^{-1}a^{1/2}(c_1\sqrt{\frac{\log d}{n}}+b)$. This shows that in general the estimator is inconsistent for d fixed and $n\to\infty$. When $b=0$, we recover Theorem 1 of [10], with a slight improvement on the requirement on the sample size. Here the condition on n reads $n\gtrsim\tau^{-2}a^2\log(d)$, whereas Theorem 1 of [10] imposes $n\gtrsim\tau^{-2}a^3\log(d)$.

Although the estimator is inconsistent, if b is small, $\hat{\theta}_n$ would still give a reasonably good estimate of $\bar{\theta}$. In such cases, if in addition $\min_{(s,\ell)\in I}|\theta_\star(s,\ell)|$ is comparatively large, one can also correctly recover the sign of $\{\theta_\star(s,\ell),\ (s,\ell)\in I\}$ by a simple hard-thresholding rule, where the sign of a vector is defined as the vector of signs. Consider the estimator $\tilde{\theta}_n$ where

$$\tilde{\theta}_n(s,\ell) = \begin{cases}\hat{\theta}_n(s,\ell) & \text{if } |\hat{\theta}_n(s,\ell)|>\delta\\ 0 & \text{otherwise,}\end{cases}$$

for a thresholding parameter δ. Following Corollary 2 of [14], the next result is a direct consequence of Theorem 1.2.

Corollary 1.3. *Under the assumptions of Theorem 1.2, and assuming that $\delta>26\sqrt{a}\lambda_n$, and $\min_{(s,\ell)\in I}|\theta_\star(s,\ell)|>26\sqrt{a}\lambda_n$,*

$$\mathbb{P}_\star\left[\mathsf{sign}(\tilde{\theta}_n)=\mathsf{sign}(\bar{\theta}_\star)\right]\geq 1-\frac{3}{d}.$$

1.2.1. On the misspecification parameter b

The misspecification parameter b plays an important role in the results above. Clearly b is related to how much connections there is between the observed and the missing nodes. However, as the next result shows, b is controlled mainly by the strength of the connections between the observed and the missing nodes, not necessarily by the number of missing nodes.

Proposition 1.4.

$$b \leq \left\{ \sup_{x,y} |B(x,y)| \right\} \wedge \left\{ c_1^2 \sup_{s \in \mathcal{D}} \sum_{j \in \partial_s \setminus \mathcal{D}} |\theta_\star(s,j)| \right\}, \tag{1.9}$$

where $a \wedge b \overset{\text{def}}{=} \min(a,b)$.

Proof. The fact b is smaller than $\sup_{x,y} |B(x,y)|$ follows directly from its definition. Recall that $\mathbb{E}_\star$ denotes the expectation under the true model $\theta_\star$. Hence, by first conditioning on $\{X_\ell,\ \ell \in \mathcal{S} \setminus \{s\}\}$, we have

$$\mathbb{E}_\star \left[B(X_s, X_\ell) \right] = \mathbb{E}_\star \left[\int B(u, X_\ell) f_{\theta_\star}^{(s)}(u|X_{\partial_s}) du \right],\ s \neq \ell.$$

Hence

$$\begin{aligned} &\mathbb{E}_\star \left[B(X_s, X_\ell) \right] - \mathbb{E}_\star \left[\int B(u, X_\ell) \bar{f}_{\theta_\star}^{(s)}(u|X_{\mathcal{D}_s}) du \right] \\ &\qquad = \mathbb{E}_\star \left[\int B(u, X_\ell) f_{\theta_\star}^{(s)}(u|X_{\mathcal{S}_s}) du - \int B(u, X_\ell) \bar{f}_{\theta_\star}^{(s)}(u|X_{\mathcal{D}_s}) du \right]. \end{aligned}$$

We then apply (A.3), with $f_1 = f_2 = B(\cdot, X_\ell)$ which gives

$$\left| \int B(u, X_\ell) f_{\theta_\star}^{(s)}(u|X_{\mathcal{S}_s}) du - \int B(u, X_\ell) \bar{f}_{\theta_\star}^{(s)}(u|X_{\mathcal{D}_s}) du \right| \leq c_1^2 \sum_{j \in \partial_s \setminus \mathcal{D}} |\theta_\star(s,j)|,$$

and the stated result follows easily. □

Better bounds than (1.9) can be derived with additional assumptions. And in turn, such bounds can be used to state Theorem 1.2 with different assumptions. Consider for example the case of the Ising model where

$$\mathsf{X} = \{0,1\}, \quad \text{and} \quad B(x,y) = xy. \tag{1.10}$$

For $s \in \mathcal{D}$, we define

$$\mathsf{deg}_{\mathsf{out}}(s) \overset{\text{def}}{=} \sum_{j \in \mathcal{S} \setminus \mathcal{D}} \theta_\star(s,j), \quad \text{and} \quad \mathsf{deg}_{\mathsf{in}}(s) \overset{\text{def}}{=} \sum_{j \in \mathcal{D}_s} \theta_\star(s,j),$$

as the strength of the connection between s and the missing nodes, and between s and the observed nodes, respectively. We will also assume that there are only non-negative interactions in the network: $\theta_\star(s,\ell) \geq 0$ for all $(s,\ell) \in \underline{\mathcal{S}}^2$. This assumption is mostly technical and made to simplify the analysis. Recall that c_1 is defined in (1.3), $\mathcal{D} \subseteq \mathcal{S}$, $p = |\mathcal{D}|$, and $d = p(p-1)/2$.

Corollary 1.5. *Assume A1, (1.10), and suppose that $\theta_\star(s,\ell) \geq 0$ for all $(s,\ell) \in \underline{\mathcal{S}}^2$. Take $\lambda_n = 10c_1\sqrt{\frac{\log d}{n}}$. If n, $\mathcal{D}$, $\mathcal{S}$, and $\theta_\star$ are such that $n\tau^2 \geq 2(64^2)c_1^2a^2\log(2d)$, $(480)c_1a\sqrt{\frac{\log d}{n}} < \tau$, and*

$$\sup_{s\in\mathcal{D}}\left\{\mathsf{deg}_{\mathit{out}}(s)\exp\left(-\frac{1}{2}\mathsf{deg}_{\mathit{in}}(s)\right)\right\} \leq \frac{c_1}{2}\sqrt{\frac{\log d}{n}}. \tag{1.11}$$

Then

$$\left\|\hat{\theta}_n - \bar{\theta}_\star\right\|_2 \leq (260)c_1\sqrt{\frac{a\log d}{n}},$$

with a probability at least $1 - \frac{3}{d}$

Remark 2. The result formalizes the intuition that the estimation procedure will work well when $\mathcal{D}$ is large and the nodes in $\mathcal{D}$ interact only weakly with the missing nodes. If $\mathsf{deg}_{\mathsf{in}}(s)$ is large, $\mathsf{deg}_{\mathsf{out}}(s)$ would need to be exponentially larger to result in a large bias. However, note that the result still does not imply that $\hat{\theta}_n$ is consistent for d fixed, since for d fixed, (1.11) will fail as $n \to \infty$.

Obviously this result is not of much practical use, because the $\mathsf{deg}_{\mathsf{in}}(s)$ and $\mathsf{deg}_{\mathsf{out}}(s)$ are typically unknown. A more promising approach would be to develop misspecification statistical testings to detect the existence of the missing nodes. However this is beyond the scope of the paper, and is left for possible future research.

Proof. It suffices to show that

$$b \leq \sup_{s\in\mathcal{D}}\left\{\mathsf{deg}_{\mathsf{out}}(s)\exp\left(-\frac{1}{2}\mathsf{deg}_{\mathsf{in}}(s)\right)\right\}, \tag{1.12}$$

which together with (1.11) implies that $\lambda_n = 10c_1\sqrt{\frac{\log d}{n}} \geq 4b + 8c_1\sqrt{\frac{\log d}{n}}$, and $48c_1a\lambda_n = 480c_1a\sqrt{\frac{\log d}{n}} < \tau$. The corollary then follows from Theorem 1.2.

It remains to show (1.12). We denote $\mathsf{logit}^{-1}(x) = \frac{e^x}{1+e^x}$, and $G(x) = e^x(1+e^x)^{-2}$ its derivative. In the case of the Ising model, the conditional means are given by $\mathbb{E}_\star[X_s|X_{\mathcal{S}\setminus\{s\}}] = \mathsf{logit}^{-1}(\sum_{j\in\partial_s}\theta_\star(s,j)X_j)$. Hence

$$\mathbb{E}_\star\left[B(X_s,X_\ell)|X_{\mathcal{S}\setminus\{s\}}\right] = X_\ell\mathbb{E}_\star\left[X_s|X_{\mathcal{S}\setminus\{s\}}\right] = X_\ell\mathsf{logit}^{-1}\left(\sum_{j\in\partial_s}\theta_\star(s,j)X_j\right).$$

Similarly

$$\int_{\mathsf{X}} B(u,X_\ell)\bar{f}^{(s)}_{\theta_\star}(u|X_{\mathcal{D}_s})du = X_\ell\mathsf{logit}^{-1}\left(\sum_{j\in\partial_s\cap\mathcal{D}}\theta_\star(s,j)X_j\right).$$

Notice that $G(x) \leq e^{-|x|}$, for all $x \in \mathbb{R}$. Hence, by Taylor expansion

$$
\begin{aligned}
& X_\ell \left(\mathsf{logit}^{-1} \left(\sum_{j \in \partial s} \theta_\star(s,j) X_j \right) - \mathsf{logit}^{-1} \left(\sum_{j \in \partial_s \cap \mathcal{D}} \theta_\star(s,j) X_j \right) \right) \\
& = X_\ell \left\{ \sum_{j \in \partial_s \setminus \mathcal{D}} \theta_\star(s,j) X_j \right\} \int_0^1 G \left(\sum_{j \in \partial_s \cap \mathcal{D}} \theta_\star(s,j) X_j + t \sum_{j \in \partial_s \setminus \mathcal{D}} \theta_\star(s,j) X_j \right) dt \\
& \leq \left\{ \sum_{j \in \partial_s \setminus \mathcal{D}} \theta_\star(s,j) X_j \right\} \int_0^1 \exp \left(- \sum_{j \in \partial_s \cap \mathcal{D}} \theta_\star(s,j) X_j - t \sum_{j \in \partial_s \setminus \mathcal{D}} \theta_\star(s,j) X_j \right) dt.
\end{aligned}
$$

For all $x, y \geq 0$, it is clear that $y \int_0^1 e^{-x-ty} dt = e^{-x} \int_0^1 y e^{-ty} dt = e^{-x}(1 - e^{-y})$, which easily yields the bound

$$
b \leq \sup_{s \in \mathcal{D}} \mathbb{E}_\star \left[\exp \left(- \sum_{j \in \mathcal{D}_s} \theta_\star(s,j) X_j \right) \left\{ 1 - \exp \left(- \sum_{j \in \mathcal{S} \setminus \mathcal{D}} \theta_\star(s,j) X_j \right) \right\} \right].
$$

Since $1 - e^{-y} \leq y$ for all $y \geq 0$, and using also the Jensen's inequality, we have

$$
b \leq \sup_{s \in \mathcal{D}} \left\{ \sum_{j \in \mathcal{S} \setminus \mathcal{D}} \theta_\star(s,j) \right\} \exp \left(- \sum_{j \in \mathcal{D}_s} \theta_\star(s,j) \mathbb{E}_\star(X_j) \right).
$$

We saw earlier that $\mathbb{E}_\star[X_s | X_{\mathcal{S} \setminus \{s\}}] = \mathsf{logit}^{-1}(\sum_{j \in \partial_s} \theta_\star(s,j) X_j) \geq \frac{1}{2}$, and (1.12) follows. □

1.2.2. On assumption A1

A1 is a type of restricted eigenvalue assumption similar to the Assumption $RE(s, c_0)$ of [5]. This assumption is not easy to check. But following the analysis of [5], it is possible to derive sufficient conditions that give some intuition into when A1 holds. For simplicity we consider the case of product-form functions: $B(x,y) = B_0(x) B_0(y)$. Then

$$
\begin{aligned}
& H^{(s)}(\ell, \ell'; X) \\
& = B_0(X_\ell) B_0(X_{\ell'}) \left[\int_{\mathsf{X}} B(u)^2 \bar{f}^{(s)}_{\theta_\star}(u | X_{\mathcal{D}_s}) du - \left\{ \int_{\mathsf{X}} B(u) \bar{f}^{(s)}_{\theta_\star}(u | X_{\mathcal{D}_s}) du \right\}^2 \right].
\end{aligned}
$$

Thus assuming that there exists a finite constant $\alpha > 0$ such that

$$
\min_{s \in \mathcal{D}} \left\{ \int_{\mathsf{X}} B(u)^2 \bar{f}^{(s)}_{\theta_\star}(u | X_{\mathcal{D}_s}) du - \left\{ \int_{\mathsf{X}} B(u) \bar{f}^{(s)}_{\theta_\star}(u | X_{\mathcal{D}_s}) du \right\}^2 \right\} \geq \alpha, \tag{1.13}
$$

we have

$$\theta_s' C^{(s)} \theta_s \geq \alpha \sum_{\ell \in \mathcal{D}_s} \sum_{\ell' \in \mathcal{D}_s} \theta(s,\ell)\theta(s,\ell') \mathbb{E}_\star \left(B_0(X_\ell) B_0(X_{\ell'}) \right)$$

$$= \alpha \mathbb{E}_\star \left[\left(\sum_{\ell \in \mathcal{D}_s} \theta(s,\ell) B_0(X_\ell) \right)^2 \right] \geq \alpha \mathsf{Var}_\star \left(\sum_{\ell \in \mathcal{D}_s} \theta(s,\ell) B_0(X_\ell) \right). \quad (1.14)$$

Assumption (1.13) is similar to Assumption 2 of [13], and is typically not restrictive. We show below that it holds for the auto-logistic model. The difficulty lies in dealing with the covariance matrix of the (local) fields $\{B_0(X_\ell),\ \ell \in \mathcal{D}_s\}$. Following [5] (Section 4), the next result captures the intuition that the covariance matrix of $\{B_0(X_\ell),\ \ell \in \mathcal{D}_s\}$ is positive definite if the covariance matrix between the neighbors of s is positive definite and the covariance between neighbors of s and non-neighbors of s is weak. This bears some similarity with the dependency assumption A and the incoherence assumption B of [10].

Proposition 1.6. *Assume (1.13) and suppose that for all $u \in \mathbb{R}^{|\partial s|}$,*

$$\inf_{s \in \mathcal{D}} \sum_{\ell \in \partial s} \sum_{\ell' \in \partial s} u_\ell u_{\ell'} \mathsf{Cov}_\star \left(B_0(X_\ell), B_0(X_{\ell'}) \right) \geq \rho \sum_{\ell \in \partial s} u_\ell^2, \quad \textit{and}$$

$$\sup_{s \in \mathcal{D}} \sup_{j \notin \partial s} \sum_{\ell \in \partial s} u_\ell \mathsf{Cov}_\star \left(B_0(X_\ell), B_0(X_j) \right) \leq \delta \sqrt{\sum_{\ell \in \partial s} u_\ell^2}.$$

Then for all $\theta \in \Delta$,

$$\sum_{s \in \mathcal{D}} \theta_s' C^{(s)} \theta_s \geq 2\alpha \left(\rho - 6a^{1/2}\delta \right) \|\theta\|_{2\star}^2.$$

Proof. We have

$$\mathsf{Var}_\star \left(\sum_{\ell \in \mathcal{D}_s} \theta(s,\ell) B_0(X_\ell) \right) \geq \mathsf{Var}_\star \left(\sum_{\ell \in \partial s \cap \mathcal{D}} \theta(s,\ell) B_0(X_\ell) \right)$$

$$+ 2 \sum_{\ell \in \partial s \cap \mathcal{D}} \sum_{\ell' \in \mathcal{D}_s \setminus \partial s} \theta(s,\ell)\theta(s,\ell') \mathsf{Cov}_\star \left(B_0(X_\ell), B_0(X_{\ell'}) \right)$$

$$\geq \rho \sum_{\ell \in \partial s \cap \mathcal{D}} \theta(s,\ell)^2 - 2\delta \sqrt{\sum_{\ell \in \partial s \cap \mathcal{D}} |\theta(s,\ell)|^2} \sum_{\ell' \in \mathcal{D}_s \setminus \partial s} |\theta(s,\ell')|$$

$$\geq \rho \sum_{\ell \in \partial s \cap \mathcal{D}} \theta(s,\ell)^2 - 2\delta \|\theta\|_{2\star} \sum_{\ell' \in \mathcal{D}_s \setminus \partial s} |\theta(s,\ell')|$$

Clearly $\sum_{s \in \mathcal{D}} \sum_{\ell \in \partial s \cap \mathcal{D}} \theta(s,\ell)^2 = 2\|\theta\|_{2\star}^2$, and

$$\sum_{s \in \mathcal{D}} \sum_{\ell' \in \mathcal{D}_s \setminus \partial s} |\theta(s,\ell')| = 2 \sum_{(s,\ell) \in \underline{\mathcal{D}}^2 \setminus I} |\theta(s,\ell)| \leq 6 \sum_{(s,\ell) \in I} |\theta(s,\ell)| \leq 6a^{1/2} \|\theta\|_{2\star}.$$

Therefore, using (1.14), we get

$$\sum_{s\in\mathcal{D}} \theta_s' C^{(s)}\theta_s \geq \alpha \left(2\rho\|\theta\|_{2\star}^2 - 12a^{1/2}\delta\|\theta\|_{2\star}^2\right) \geq 2\alpha\left(\rho - 6a^{1/2}\delta\right)\|\theta\|_{2\star}^2.$$

□

1.3. Example and Monte Carlo evidence

We consider the example of the auto-logistic model where $\mathsf{X} = \{0,1\}$, and $B(x,y) = xy$. For the simulations, we consider three cases: $p = 50$, $p = 80$, and $p = 100$. For each setting, we consider different sample sizes $n \in \{50, 100, \ldots, 500\}$. The sparsity (the proportion of non-zero entries) of $\theta_\star$ is set to 1%, and the non-zero entries of $\theta_\star$ are generated uniformly from the interval $[0.3, 1]$. Having all the entries of $\theta_\star$ non-negative allows us to simulate exactly (instead of using MCMC) from the Ising distribution $\mu_{\theta_\star}$ using the Propp-Wilson's perfect sampler. For all the simulations, we set the regularization parameter to $\lambda = 0.5\sqrt{\log p)/n}$.

To quantify the amount of missing data, we use the upper bound established above

$$\check{b} \overset{\text{def}}{=} \sup_{s\in\mathcal{D}} \sum_{\ell\in\partial_s\setminus\mathcal{D}} |\theta_\star(s,\ell)| \exp\left(-\frac{1}{2}\sum_{j\in\mathcal{D}_s}\theta_\star(s,j)\right).$$

We compare three settings. In Setting 1, there is no missing data, and the samples are generated exactly from (1.1). In Setting 2 and 3, we generate the sample $(X_1^{(i)},\ldots,X_p^{(i)},X_{p+1}^{(i)},\ldots,X_{p+r}^{(i)})$ from (1.1), for $\theta = \theta_\star$, and we retain only $(X_1^{(i)},\ldots,X_p^{(i)})$, for $1 \leq i \leq n$. Thus there are r missing nodes. In Setting 2, we use $r = 2$, whereas in Setting 3, we set $r = 20$. Table 1 shows the corresponding values of $\check{b}$ in each setting.

Regardless of the data generation mechanism, we fit model (1.1) by ℓ^1 penalized pseudo-likelihood and compute the relative Mean Square Error $\mathbb{E}_\star(\|\hat{\theta}_n - \bar{\theta}_\star\|_2)/\|\bar{\theta}_\star\|_2$, estimated from K replications of the estimator ($K = 10$). In Figure 1, we plot $\mathbb{E}_\star(\|\hat{\theta}_n - \bar{\theta}_\star\|_2)/\|\bar{\theta}_\star\|_2$ as a function of the sample size.

As expected, the estimation error decreases with the sample size. Also, the more missing data, the worst the estimator behaves. We also observe that in Setting 2 where $r = 2$, the value of $\check{b}$ is the same in the cases $p = 50$ and $p = 100$, but the estimation error is noticeably more affected by the missing data for $p = 50$. This seems in agrement with the conclusion of Corollary 1.5.

Table 1
Values of $\check{b}$ in each setting of the simulation

	Setting 1, $r = 0$	Setting 2, $r = 2$	Setting 3, $r = 20$
$p = 50$	0	1.8	2.3
$p = 80$	0	1.0	2.1
$p = 100$	0	1.8	2.9

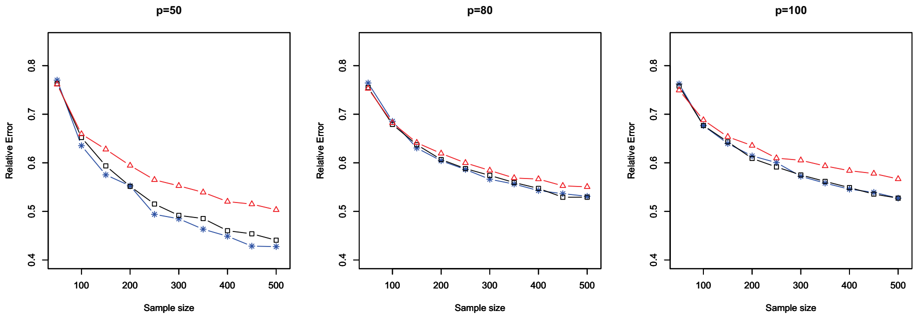

FIG 1. *Relative MSE versus sample size n, where star-line is Setting 1, square-line is Setting 2, triangle-line is Setting 3.*

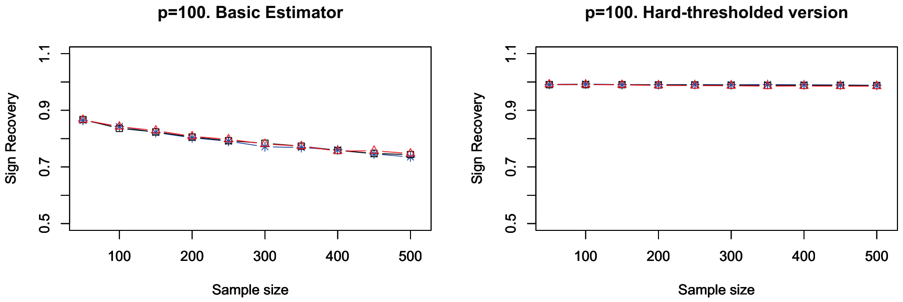

FIG 2. *Proportion of sign correctly recovered versus sample size n, where star-line is Setting 1, square-line is Setting 2, triangle-line is Setting 3.*

We also compute the proportion of signs of $\theta_\star$ that is correctely recovered. This is plotted in Figure 2 for $p = 100$. The estimator $\tilde{\theta}_n$ described in Corollary 1.3 performs much better than the initial estimator $\hat{\theta}_n$. For the computation of $\tilde{\theta}_n$, we follow Corollary 1.3 and use a threshold of $\delta = \sqrt{a}\lambda$, where a is the number of non-zero entries of $\theta_\star$. For large sample size, the sign recovery of $\tilde{\theta}_n$ is perfect, even in presence of missing nodes (the three lines are almost undistinguishable). Notice that the penalty parameter $\lambda = 0.5\sqrt{\log p/n}$ varies (decreases) with n. This negatively affects the basic estimator $\hat{\theta}_n$ but does not affect $\tilde{\theta}_n$.

2. Proof of Theorem 1.2

We define

$$U_n(\theta) = -Q_n(\bar{\theta}_\star + \theta) + Q_n(\bar{\theta}_\star), \quad \theta \in \mathcal{M}(\mathcal{D}).$$

Clearly, U_n is strictly convex, $U_n(0) = 0$, and minimized at $\hat{\theta}_n - \bar{\theta}_\star$. We recall that

$$\Delta = \left\{ \theta \in \mathcal{M}(\mathcal{D}) : \sum_{(s,\ell)\in\underline{\mathcal{D}}^2\setminus I} |\theta(s,\ell)| \leq 3 \sum_{(s,\ell)\in I} |\theta(s,\ell)| \right\},$$

and for $r > 0$ we set

$$\Delta_r \overset{\text{def}}{=} \{\theta \in \Delta : \ \|\theta\|_2 = r\} .$$

The next two lemmas are adaptations of Lemmas 1 and 4 from [15]. We give a proof for completeness.

Lemma 2.1. *On the event $\{\|\nabla \ell_n(\bar{\theta}_\star)\|_\infty \leq \frac{\lambda_n}{2}\}$, $(\hat{\theta}_n - \bar{\theta}_\star) \in \Delta$.*

Proof. $-\ell_n(\theta) = -\sum_{i=1}^n \sum_{s\in\mathcal{D}} \log f_\theta^{(s)}(X_s^{(i)}|X_{\mathcal{D}_s}^{(i)})$, which is a convex function of θ by virtue of Lemma A.1. It is also not hard to see that

$$\|\bar{\theta}_\star + \theta\|_1 \geq \|\bar{\theta}_\star\|_1 + \sum_{(s,\ell)\notin I} |\theta(s,\ell)| - \sum_{(s,\ell)\in I} |\theta(s,\ell)|. \tag{2.1}$$

Therefore, using the convexity of $-\ell_n$ and (2.1), it follows that on $\{\|\nabla \ell_n(\theta_\star)\|_\infty \leq \lambda_n/2\}$,

$$\begin{aligned}
U_n(\theta) &= \left(-\ell_n(\bar{\theta}_\star + \theta) + \ell_n(\bar{\theta}_\star)\right) + \lambda_n \left(\|\bar{\theta}_\star + \theta\|_1 - \|\bar{\theta}_\star\|_1\right) \\
&\geq \left\langle -\nabla \ell_n(\bar{\theta}_\star), \theta \right\rangle + \lambda_n \left(\sum_{(s,\ell)\notin I} |\theta(s,\ell)| - \sum_{(s,\ell)\in I} |\theta(s,\ell)| \right) \\
&\geq -\frac{\lambda_n}{2}\|\theta\|_1 + \lambda_n \left(\sum_{(s,\ell)\notin I} |\theta(s,\ell)| - \sum_{(s,\ell)\in I} |\theta(s,\ell)| \right) \\
&= \lambda_n \left(\frac{1}{2} \sum_{(s,\ell)\notin I} |\theta(s,\ell)| - \frac{3}{2} \sum_{(s,\ell)\in I} |\theta(s,\ell)| \right).
\end{aligned}$$

Since, $U_n(0) = 0$, we necessarily have $U_n(\hat{\theta}_n - \bar{\theta}_\star) \leq 0$, which implies, in view of the above bound, that $\hat{\theta}_n - \bar{\theta}_\star \in \Delta$. □

Lemma 2.2. *On the event $\{\inf_{v\in\Delta_r} U_n(v) > 0, \ \ \text{and} \ \ \|\nabla \ell_n(\bar{\theta}_\star)\|_\infty \leq \frac{\lambda_n}{2}\}$, $\|\hat{\theta}_n - \bar{\theta}_\star\| \leq r$.*

Proof. Suppose that $\|\hat{\theta}_n - \bar{\theta}_\star\| > r$, and that $\{\|\nabla \ell_n(\bar{\theta}_\star)\|_\infty \leq \lambda_n/2\}$ occurs. We will show that there exists $\vartheta \in \Delta_r$ such that $U_n(\vartheta) \leq 0$, and this proves the result.

By Lemma 2.1, on $\{\|\nabla \ell_n(\bar{\theta}_\star)\|_\infty \leq \lambda_n/2\}$, $(\hat{\theta}_n - \bar{\theta}_\star) \in \Delta$. Assuming that $\|\hat{\theta}_n - \bar{\theta}_\star\| > r$, we can find $\alpha \in (0,1)$ such that $\alpha\|\hat{\theta}_n - \bar{\theta}_\star\| = r$. It is also clear that if $\theta \in \Delta$, $t\theta \in \Delta$ for all $t \geq 0$. Hence $\alpha(\hat{\theta}_n - \bar{\theta}_\star) \in \Delta_r$. But by convexity

$$U_n\left(\alpha(\hat{\theta}_n - \bar{\theta}_\star)\right) = U_n\left(\alpha(\hat{\theta}_n - \bar{\theta}_\star) + (1-\alpha)0\right) \leq \alpha U_n\left(\hat{\theta}_n - \bar{\theta}_\star\right) \leq 0.$$

□

The main idea of the proof is to show that under the assumption of the theorem the event $\{\inf_{v\in\Delta_r} U_n(v) > 0, \ \ \text{and} \ \ \|\nabla \ell_n(\bar{\theta}_\star)\|_\infty \leq \frac{\lambda_n}{2}\}$ occurs with high probability with $r = r_n$ appropriately chosen. To make the proof easier to follow, we include the following intermediary step. For $s \in \mathcal{D}$, $\vartheta \in \mathbb{R}^{p-1}$, and $x \in \mathsf{X}^{\mathcal{D}_s}$, we define

$$V^{(s)}(\vartheta, x) \overset{\text{def}}{=} \int \left(\sum_{\ell \in \mathcal{D}_s} \vartheta_\ell B(u, x_\ell) \right)^2 \bar{f}^{(s)}_{\theta_\star}(u|x) du \\ - \left(\int \sum_{\ell \in \mathcal{D}_s} \vartheta_\ell B(u, X_\ell) \bar{f}^{(s)}_{\theta_\star}(u|x) du \right)^2 . \tag{2.2}$$

We recall the notation $\theta_s = \{\theta(s, \ell),\ \ell \in \mathcal{D}_s\}$.

Lemma 2.3. *Consider the event*

$$\mathcal{E}_n(\tau) \overset{\text{def}}{=} \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{s \in \mathcal{D}} V^{(s)}(\theta_s, X^{(i)}) \geq \tau \|\theta\|^2_{2\star} \ \textit{for all}\ \ \theta \in \Delta, \\ \textit{and} \quad \left\| \nabla \ell_n(\bar{\theta}_\star) \right\|_\infty \leq \frac{\lambda_n}{2} \right\}.$$

Suppose that there exists $\tau > 0$ such that $\tau > 48 c_1 a \lambda_n$, and the event $\mathcal{E}_n(\tau)$ holds. Then

$$\|\hat{\theta}_n - \bar{\theta}_\star\|_2 \leq 26 a^{1/2} \lambda_n.$$

Proof. We know from Lemma 2.1 that on $\mathcal{E}_n(\tau)$, $\hat{\theta}_n - \bar{\theta}_\star \in \Delta$. Set $r_n = 26 a^{1/2} \lambda_n$. We will show that on $\mathcal{E}_n(\tau)$, $\inf_{\theta \in \Delta_{r_n}} U_n(\theta) > 0$, and use Lemma 2.2 to conclude that $\|\hat{\theta}_n - \bar{\theta}_\star\|_2 \leq r_n$. We recall that for $\theta \in \mathcal{M}(\mathcal{D})$,

$$U_n(\theta) = \left(-\ell_n(\bar{\theta}_\star + \theta) + \ell_n(\bar{\theta}_\star) + \left\langle \nabla \ell_n(\bar{\theta}_\star), \theta \right\rangle \right) - \left\langle \nabla \ell_n(\bar{\theta}_\star), \theta \right\rangle \\ + \lambda_n \left(\|\bar{\theta}_\star + \theta\|_1 - \|\theta_\star\|_1 \right).$$

For $\theta \in \Delta$, and on the event $\left\{ \|\nabla \ell_n(\bar{\theta}_\star)\|_\infty \leq \lambda_n/2 \right\}$

$$|- \langle \nabla \ell_n(\theta_\star), \theta \rangle + \lambda_n \left(\|\theta_\star + \theta\|_1 - \|\theta_\star\|_1 \right)| \leq 6 \lambda_n a^{1/2} \|\theta\|_{2\star}. \tag{2.3}$$

From the expression of the approximate conditional distribution $\bar{f}^{(s)}_\theta(x_s|x)$ given in (1.5), we have

$$- \ell_n(\bar{\theta}_\star + \theta) + \ell_n(\bar{\theta}_\star) + \left\langle \nabla \ell_n(\bar{\theta}_\star), \theta \right\rangle \\ = \sum_{s \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^n \left(\log \bar{Z}^{(s)}_{\bar{\theta}_\star + \theta}(X^{(i)}) - \log \bar{Z}_{\bar{\theta}_\star}(X^{(i)}) - \left\langle \nabla_\theta \log \bar{Z}^{(s)}_{\bar{\theta}_\star}(X^{(i)}), \theta_s \right\rangle \right), \tag{2.4}$$

where $\bar{Z}^{(s)}_\theta(x) = \int \exp(\sum_{\ell \in \mathcal{D}_s} \theta(s, \ell) B(u, x_\ell)) du$, $\theta_s = \{\theta(s, \ell),\ \ell \in \mathcal{D}_s\}$, and $\langle u, v \rangle$ here denotes the usual inner product in $\mathbb{R}^{\mathcal{D}_s}$. Fix $s \in \mathcal{D}$, $x \in \mathsf{X}^{\mathcal{D}_s}$. We will now apply the self-concordant bound developed in Lemma A.2 to the function

$$\zeta^{(s)}(\vartheta|x) \stackrel{\text{def}}{=} \int_{\mathsf{X}} \exp\left(\sum_{\ell\in\mathcal{D}_s} \vartheta_\ell B(u,x_\ell)\right) du.$$

The constant c is Lemma A.2 is given here by $\sup_{u\in\mathsf{X}}\sup_{x,y\in\mathsf{X}} |B(u,x)-B(u,y)| = c_1$. Hence

$$\log \bar{Z}^{(s)}_{\bar{\theta}_\star+\theta}(X^{(i)}) - \log \bar{Z}_{\bar{\theta}_\star}(X^{(i)}) - \left\langle \nabla_\theta \log \bar{Z}^{(s)}_{\bar{\theta}_\star}(X^{(i)}), \theta_s \right\rangle \\ \geq \frac{V^{(s)}(\theta_s, X^{(i)}_{-s})}{2 + c_1 \sum_{\ell\in\mathcal{D}_s} |\theta(s,\ell)|}. \quad (2.5)$$

Since $\sum_{\ell\in\mathcal{D}_s} |\theta(s,\ell)| \leq \|\theta\|_1$, we combine the above with (2.4) to conclude that

$$-\ell_n(\bar{\theta}_\star+\theta) + \ell_n(\bar{\theta}_\star) + \left\langle \nabla\ell_n(\bar{\theta}_\star), \theta\right\rangle \\ \geq \frac{1}{2+c_1\|\theta\|_1} \frac{1}{n}\sum_{i=1}^n \sum_{s\in\mathcal{D}} V^{(s)}(\theta_s, X^{(i)}) \geq \frac{\tau\|\theta\|^2_{2\star}}{2+c_1\|\theta\|_1}, \quad (2.6)$$

for all $\theta\in\Delta$, using the fact that $\mathcal{E}_n(\tau)$ holds. This bound and (2.3) yield that for $\theta\in\Delta$,

$$U_n(\theta) \geq \frac{\tau\|\theta\|^2_{2\star}}{2+4c_1 a^{1/2}\|\theta\|_{2\star}} - 6\lambda_n a^{1/2}\|\theta\|_{2_\star}.$$

The right-hand-side is positive whenever

$$\|\theta\|_{2\star} > \frac{13a^{1/2}\lambda_n}{\tau - 24c_1 a\lambda_n} \geq 26a^{1/2}\lambda_n,$$

provided $48c_1 a\lambda_n < \tau$. □

We now show that the event $\mathcal{E}_n(\tau)$ occurs with high probability.

Lemma 2.4. *For any* $\lambda_n \geq 4b + 8c_1\sqrt{\frac{\log d}{n}}$,

$$\mathbb{P}_\star\left[\|\nabla\ell_n(\theta_\star)\|_\infty > \frac{\lambda_n}{2}\right] \leq \frac{2}{d}. \quad (2.7)$$

Proof. Set $\delta_n = 8c_1\sqrt{\frac{\log d}{n}}$. We calculate that for $(s,\ell)\in\underline{\mathcal{D}}^2$,

$$\frac{\partial\ell_n(\theta_\star)}{\partial\theta_{s,\ell}} = \frac{1}{n}\sum_{i=1}^n 2B(X^{(i)}_s, X^{(i)}_\ell) - \int B(u, X^{(i)}_\ell)\bar{f}^{(s)}_{\theta_\star}(u|X^{(i)}_{\mathcal{D}_s})du \\ - \int B(u, X^{(i)}_s) f^{(\ell)}_{\theta_\star}(u|X^{(i)}_{\mathcal{D}_\ell})du.$$

By the definition of b,

$$\left|\mathbb{E}_\star\left[2B(X_s^{(i)}, X_\ell^{(i)}) - \int B(u, X_\ell^{(i)})\bar{f}_{\theta_\star}^{(s)}(u|X_{\mathcal{D}_s}^{(i)})du \right.\right.$$
$$\left.\left. - \int B(u, X_s^{(i)})f_{\theta_\star}^{(\ell)}(u|X_{\mathcal{D}_\ell}^{(i)})du\right]\right| \leq 2b.$$

Therefore for each $(s,\ell) \in \underline{\mathcal{D}}^2$, and by Hoeffding's inequality

$$\begin{aligned} \mathbb{P}_\star\left[\left|\frac{\partial \ell_n(\theta_\star)}{\partial \theta_{s,\ell}}\right| > \frac{\lambda_n}{2}\right] &\leq \mathbb{P}_\star\left[\left|\frac{\partial \ell_n(\theta_\star)}{\partial \theta_{s,\ell}} - \mathbb{E}_\star\left(\frac{\partial \ell_n(\theta_\star)}{\partial \theta_{s,\ell}}\right)\right| > \frac{\delta_n}{2}\right] \\ &\leq 2\exp\left(-\frac{n\delta_n^2}{32c_1^2}\right). \end{aligned}$$

We conclude by the union-sum inequality that

$$\mathbb{P}_\star\left(\|\nabla \ell_n(\theta_\star)\|_\infty > \frac{\lambda_n}{2}\right) \leq 2\exp\left(\log d - \frac{n\delta_n^2}{32c_1^2}\right) \leq \frac{2}{d},$$

given the choice $\delta_n = 8c_1\sqrt{\frac{\log d}{n}}$. □

Lemma 2.5. *Assume A1, and let τ as in A1. Suppose that $n\tau^2 \geq 4(64^2)c_1^4 a^2 \log(2d)$. Then*

$$\mathbb{P}_\star\left[\frac{1}{n}\sum_{i=1}^n \sum_{s\in\mathcal{D}} V^{(s)}(\theta_s, X^{(i)}) \geq \tau\|\theta\|_{2\star}^2 \;\; \textit{for all} \;\; \theta \in \Delta\right] > 1 - \frac{1}{d}.$$

Proof. Set

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^n \sum_{s\in\mathcal{D}} V^{(s)}(\theta_s, X^{(i)}).$$

We recall that the definition of $V^{(s)}$ is given in (2.2). It is worth noticing that

$$V^{(s)}(\theta_s, X^{(i)}) = \sum_{\ell\in\mathcal{D}_s}\sum_{\ell'\in\mathcal{D}_s} \theta(s,\ell)\theta(s,\ell')H^{(s)}(\ell,\ell;X^{(i)}),$$

with $H^{(s)}(\ell,\ell';X^{(i)})$ as defined in (1.7). it is clear that $\mathbb{E}_\star\left(V^{(s)}(\theta_s, X^{(1)})\right) = \theta_s' C^{(s)}\theta_s$. Hence for $\theta \in \Delta$, and using A1,

$$\mathbb{E}\left(Q_n(\theta)\right) = \sum_{s\in\mathcal{D}} \mathbb{E}\left[V^{(s)}(\theta_s, X^{(1)})\right] = \sum_{s\in\mathcal{D}} \theta_s' C^{(s)}\theta_s \geq 2\tau\|\theta\|_{2\star}^2.$$

Therefore,

$$Q_n(\theta) = \left(Q_n(\theta) - \mathbb{E}_\star\left(Q_n(\theta)\right)\right) + \mathbb{E}_\star\left(Q_n(\theta)\right) \geq \left(Q_n(\theta) - \mathbb{E}_\star\left(Q_n(\theta)\right)\right) + 2\tau\|\theta\|_{2\star}^2.$$

We conclude that if there exists $\theta \in \Delta$ such that $Q_n(\theta) \leq \tau\|\theta\|_{2\star}^2$, then $|Q_n(\theta) - \mathbb{E}_\star(Q_n(\theta))| \geq \tau\|\theta\|_{2\star}^2$. Set $W_{s,\ell,\ell'}^{(i)} = H^{(s)}(\ell,\ell;X^{(i)})$. It is easy to see that

$$\begin{aligned}
&|Q_n(\theta) - \mathbb{E}_\star\left(Q_n(\theta)\right)| \\
&\quad\leq \sup_{s\in\mathcal{D}}\sup_{\ell,\ell'\in\mathcal{D}_s}\left|\frac{1}{n}\sum_{i=1}^n\left(V_{s,\ell,\ell'}^{(i)} - \mathbb{E}_\star\left(V_{s,\ell,\ell'}^{(i)}\right)\right)\right|\sum_{s\in\mathcal{D}}\left(\sum_{\ell\in\mathcal{D}_s}|\theta(s,\ell)|\right)^2 \\
&\qquad\leq 4\|\theta\|_1^2\sup_{s\in\mathcal{D}}\sup_{\ell,\ell'\in\mathcal{D}_s}\left|\frac{1}{n}\sum_{i=1}^n\left(V_{s,\ell,\ell'}^{(i)} - \mathbb{E}_\star\left(V_{s,\ell,\ell'}^{(i)}\right)\right)\right| \\
&\qquad\quad\leq 64a\|\theta\|_{2,\star}^2\sup_{s\in\mathcal{D}}\sup_{\ell,\ell'\in\mathcal{D}_s}\left|\frac{1}{n}\sum_{i=1}^n\left(V_{s,\ell,\ell'}^{(i)} - \mathbb{E}_\star\left(V_{s,\ell,\ell'}^{(i)}\right)\right)\right|.
\end{aligned}$$

Therefore, if there exists a non-zero $\theta \in \Delta$ such that $Q_n(\theta) \leq \tau\|\theta\|_{2\star}^2$, then

$$\sup_{s\in\mathcal{D}}\sup_{\ell,\ell'\in\mathcal{D}_s}\left|\sum_{i=1}^n\left(V_{s,\ell,\ell'}^{(i)} - \mathbb{E}_\star\left(V_{s,\ell,\ell'}^{(i)}\right)\right)\right| \geq \frac{n\tau}{64a}.$$

By Hoeffding's inequality, the probability of this event is bounded by

$$2\exp\left(\log(2d) - \frac{2\left(\frac{n\tau}{64a}\right)^2}{4c_1^4 n}\right) = 2\exp\left(\log(2d) - \frac{n\tau^2}{2c_1^4a^2 64^2}\right) \leq 2\exp\left(-\log(2d)\right),$$

using the condition $n\tau^2 \geq 4(64^2)c_1^4a^2\log(2d)$. This proves the lemma. □

2.1. Proof of Theorem 2.3

Take $\lambda_n \geq 4b + 8c_1\sqrt{\log d/n}$. Lemma 2.4 and Lemma 2.5 show that $\mathbb{P}(\mathcal{E}_n(\tau)) \geq 1 - \frac{3}{d}$, provided $n\tau^2 > 4(64^2)c_1^2a^2\log(2d)$. Since we have also assumed $48c_1a\lambda_n < \tau$, the theorem follows from Lemma 2.3. □

Appendix

A.1. Convexity and strong convexity-type result

Let $(\mathsf{X}, \mathcal{A}, \nu)$ be a measure space, for some positive measure ν. Let $B : \mathsf{X}\times\mathbb{R}^p \to \mathbb{R}$ be such that $x \mapsto B(x,\theta)$ is measurable, and $\int e^{B(x,\theta)}\nu(dy) < \infty$ for all $\theta \in \mathbb{R}^p$. Define

$$F(\theta) \overset{\text{def}}{=} \log\int e^{B(x,\theta)}\nu(dx),\ \ \theta \in \mathbb{R}^p.$$

We gather here two key results on F. We write $|\cdot|$ (resp. $|\cdot|_1$) to denote the Euclidean norm (resp. ℓ^1-norm) of $\mathbb{R}^p$. Lemma A.2 relies on Lemma A.3 which is taken from [1] Lemma 1.

Lemma A.1. *Suppose that the function $\theta \mapsto B(x,\theta)$ is convex for ν-almost all $x \in \mathsf{X}$. Then F is convex.*

Proof. Set $Z(\theta) = \int e^{B(x,\theta)}\nu(dy)$. For $\gamma \in (0,1)$,

$$\gamma F(\theta)+(1-\gamma)F(\theta') = \log\left[Z(\theta')\left(\int_{\mathsf{X}} \exp\left(B(x,\theta) - B(x,\theta')\right)\frac{e^{B(x,\theta')}}{Z(\theta')}\nu(dx)\right)^{\gamma}\right]$$
$$\geq \log\left[\int_{\mathsf{X}} \exp\left(\gamma B(x,\theta) + (1-\gamma)B(x,\theta')\right)\nu(dx)\right] \geq F(\gamma\theta + (1-\gamma)\theta').$$

□

Lemma A.2. *Suppose that $B(x,\theta) = \langle \theta, \psi(x)\rangle$, for some bounded measurable function $\psi :\ \mathsf{X} \to \mathbb{R}^d$. Suppose also that ν is a finite measure, and set $c \stackrel{\text{def}}{=} \sup_{1\leq i\leq p}\sup_{x,y\in\mathsf{X}} |\psi_i(x) - \psi_i(y)|$. For all $\theta, u \in \mathbb{R}^p$,*

$$F(\theta + u) - F(\theta) - \langle \nabla F(\theta), u\rangle \geq \frac{\mathsf{Var}_\theta\left(B(X,u)\right)}{2 + c|u|_1}, \tag{A.1}$$

where the variance is taken under the distribution $\mu_\theta(dx) = e^{B(x,\theta)}\nu(dx)/\int_{\mathsf{X}} e^{B(x,\theta)}\nu(dx)$.

Proof. The assumption of the lemma implies that for any $\theta \in \mathbb{R}^p$, F is differentiable at θ and

$$\nabla F(\theta) = \frac{\int \psi(x)e^{\langle\theta,\psi(x)\rangle}\nu(dx)}{\int e^{\langle\theta,\psi(x)\rangle}\nu(dx)} = \mathbb{E}_\theta\left(\psi(X)\right),$$

where the expectation is taken under the probability measure μ_θ. Fix $\theta, u \in \mathbb{R}^p$, and for $t \in \mathbb{R}$, set $g(t) = F(\theta + tu) = \log \int e^{\langle\theta+tu,\psi(x)\rangle}\nu(dx)$, so that $F(\theta + u) - F(\theta) - \langle\nabla F(\theta), u\rangle = g(1) - g(0) - g'(0)$. For $t \in \mathbb{R}$, consider the probability measure on X defined by

$$m_t(dx) = \frac{e^{\langle\theta+tu,\psi(x)\rangle}\mu(dx)}{\int e^{\langle\theta+tu,\psi(x)\rangle}\mu(dx)},$$

and write $\mathbb{E}_t$ for the expectation with respect to m_t. Clearly for $t = 0$, $m_t = \mu_\theta$. Under the assumption of the lemma, g has derivatives at any order and we verify that $g'(t) = \mathbb{E}_t\left(\langle u, \psi(X)\rangle\right)$, and

$$g''(t) = \mathsf{Var}_t\left(\langle u, \psi(X)\rangle\right), \quad \text{and} \quad g'''(t) = \mathbb{E}_t\left[\left(\langle u,\psi(X)\rangle - \mathbb{E}_t\left(\langle u,\psi(X)\rangle\right)\right)^3\right].$$

Therefore

$$|g'''(t)| \leq c|u|_1 g''(t), \quad t \in \mathbb{R}.$$

Then it follows from Lemma A.3 that

$$F(\theta + u) - F(\theta) - \langle\nabla F(\theta), u\rangle \geq \frac{e^{-c|u|_1} + c|u|_1 - 1}{c^2|u|_1^2}\mathsf{Var}_0\left(B(X,u)\right).$$

Next notice that for all $x \geq 0$ we have the inequality $e^{-x} + x - 1 \geq \frac{x^2}{2+x}$. The result follows. □

Lemma A.3. *Let $g: \mathbb{R} \to \mathbb{R}$ be a 3 times differentiable function such that $|g'''(t)| \leq cg''(t)$ for all $t \in \mathbb{R}$, where c is a finite constant. Then*

$$\frac{g''(0)}{c^2}\left(e^{-ct}+ct-1\right) \leq g(t)-g(0)-g'(0)t \leq \frac{g''(0)}{c^2}\left(e^{ct}-ct-1\right), \quad t \in \mathbb{R}.$$

Proof. The proof follows essentially from Gronwall's lemma. See [1] Lemma 1 for details. □

A.2. A comparison lemma

Lemma A.4. *Let $(\mathsf{Y}, \mathcal{A}, \nu)$ be a measure space where ν is a finite measure. Let $g_1, g_2, f_1, f_2 : \mathsf{Y} \to \mathbb{R}$ be bounded measurable functions. For $i \in \{1,2\}$, define $Z_i = \int e^{g_i(y)}\nu(dy)$. For $t \in [0,1]$, let $\bar{g}_t(\cdot) = tg_2(\cdot) + (1-t)g_1(\cdot)$ and $Z_t = \int_{\mathsf{Y}} e^{\bar{g}_t(y)}\nu(dy)$. Let $\bar{f}_t : \mathsf{Y} \to \mathbb{R}$ be such that $\bar{f}_0 = f_1$ and $\bar{f}_1 = f_2$. Suppose that $\frac{d}{dt}\bar{f}_t(y)$ exists for ν-almost all $y \in \mathsf{Y}$ and $\sup_{t\in[0,1],y\in\mathsf{Y}}|\frac{d}{dt}\bar{f}_t(y)| < \infty$. Then*

$$\int f_2(y)e^{g_2(y)}Z_{g_2}^{-1}\nu(dy) - \int f_1(y)e^{g_1(y)}Z_{g_1}^{-1}\nu(dy)$$
$$= \int_0^1 dt \int_{\mathsf{Y}} \left(\frac{d}{dt}\bar{f}_t(y)\right) e^{\bar{g}_t(y)}Z_t^{-1}\nu(dy) + \int_0^1 dt \mathsf{Cov}_t\left(\bar{f}_t(X), (g_2-g_1)(X)\right), \tag{A.2}$$

where $\mathsf{Cov}_t(U_1(X), U_2(X))$ is the covariance between $U_1(X)$ and $U_2(X)$ assuming that $X \sim e^{\bar{g}_t(y)}Z_t^{-1}$.

Proof. Under the stated assumptions, the function $t \to \int_{\mathsf{Y}} \bar{f}_t(y)e^{\bar{g}_t(y)}Z_t^{-1}\nu(dy)$ is differentiable under the integral sign and we have:

$$\int f_2(y)e^{g_2(y)}Z_{g_2}^{-1}\nu(dy) - \int f_1(y)e^{g_1(y)}Z_{g_1}^{-1}\nu(dy)$$
$$= \int_0^1 \frac{d}{dt}\left(\int_{\mathsf{Y}} \bar{f}_t(y)e^{\bar{g}_t(y)}Z_t^{-1}\nu(dy)\right)dt.$$

The identity follows by carrying the differentiation under the integral sign. □

With the choice $\bar{f}_t(y) = tf_2(y) + (1-t)f_1(y)$,

$$\mathsf{Cov}_t\left(\bar{f}_t(X), (g_2-g_1)(X)\right)$$
$$= (1-t)\mathsf{Cov}_t\left(f_1(X), (g_2-g_1)(X)\right) + t\mathsf{Cov}_t\left(f_2(X), (g_2-g_1)(X)\right).$$

Hence

$$\left|\mathsf{Cov}_t\left(\bar{f}_t(X), (g_2-g_1)(X)\right)\right| \leq \mathsf{osc}(g_2-g_1)\left((1-t)\mathsf{osc}(f_1) + t\mathsf{osc}(f_2)\right),$$

where $\mathsf{osc}(f) \overset{\text{def}}{=} \sup_{x,y\in\mathsf{Y}} |f(x)-f(y)|$ is the oscillation of f. We then obtain

$$\left|\int f_2(y)e^{g_2(y)}Z_{g_2}^{-1}\nu(dy) - \int f_1(y)e^{g_1(y)}Z_{g_1}^{-1}\nu(dy)\right| \\ \leq \|f_2-f_1\|_\infty + \frac{1}{2}\mathsf{osc}(g_2-g_1)\left(\mathsf{osc}(f_1)+\mathsf{osc}(f_2)\right). \quad (A.3)$$

We will also need the following particular case. For bounded measurable function $h_1, h_2 : \mathsf{Y} \to \mathbb{R}$, we can take $f_i(y) \equiv \log\int e^{h_i(u)}\nu(du)$, $i=1,2$, $\bar{f}_t(y) \equiv \log\int e^{th_2(u)+(1-t)h_1(u)}\nu(du)$, and $g_1 = g_2$ in the lemma and get:

$$\log\int e^{h_2(y)}\nu(dy) - \log\int e^{h_1(y)}\nu(dy) = \int_0^1 dt\left(\frac{d}{dt}\bar{f}_t\right) \\ = \int_0^1\int_{\mathsf{Y}} (h_2(y)-h_1(y))\frac{e^{th_2(u)+(1-t)h_1(u)}}{\int e^{th_2(u)+(1-t)h_1(u)}\nu(du)}\nu(dy).$$

In particular,

$$\left|\log\int e^{h_2(y)}\nu(dy) - \log\int e^{h_1(y)}\nu(dy)\right| \leq \|h_2-h_1\|_\infty. \quad (A.4)$$

Acknowledgements

I'm grateful to Jian Guo, Lisa Levina, George Michailidis, and Ji Zhu for helpful discussions. This work is partly supported by NSF grant DMS 0906631.

References

[1] Bach, F. (2010). Self-concordant analysis for logistic regression. *Electron. J. Statist.* **4** 384–414. MR2645490

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

[3] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author. MR0373208

[4] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[5] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[6] Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics* **40** 1935–1967. MR3059067

[7] d'Aspremont, A., Banerjee, O. and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30** 56–66. MR2399568

[8] Drton, M. and Perlman, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **91** 591–602. MR2090624

[9] Georgii, H.-O. (1988). *Gibbs measures and phase transitions*, vol. 9 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin. MR0956646

[10] Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Joint structure estimation for categorical markov networks. Tech. rep., Univ. of Michigan.

[11] Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906. MR2505138

[12] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

[13] Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs with the lasso. *Annals of Stat.* **34** 1436–1462. MR2278363

[14] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

[15] Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27** 538–557. MR3025133

[16] Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

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

[18] Xue, L., Zou, H. and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics* **40** 1403–1429. MR3015030

[19] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824