CONSISTENCY OF ℓ_1-REGULARIZED MAXIMUM-LIKELIHOOD FOR COMPRESSIVE POISSON REGRESSION

Yen-Huan Li and Volkan Cevher

LIONS, EPFL

ABSTRACT

We consider Poisson regression with the canonical link function. This regression model is widely used in regression analysis involving count data; one important application in electrical engineering is transmission tomography. In this paper, we establish the variable selection consistency and estimation consistency of the ℓ_1-regularized maximum-likelihood estimator in this regression model, and characterize the asymptotic sample complexity that ensures consistency even under the compressive sensing setting (or the $n \ll p$ setting in high-dimensional statistics).

***Index Terms*—** Poisson regression, maximum likelihood, ℓ_1-regularization, consistency, variable selection, sample complexity, transmission tomography

1. INTRODUCTION

The Poisson regression model with the canonical link function is pervasive in regression analysis for count data [5, 16]. Transmission tomography is one famous application of this model in electrical engineering [10], and this is the main motivation of this work. In this paper, we discuss the variable selection consistency and estimation consistency of ℓ_1-regularized maximum-likelihood (ML) estimators in this regression model. We expect that our work will validate, in a theoretically sound fashion, the use of the ℓ_1-regularized ML estimator in the Poisson regression model.

Below is a summary of our main contributions.

1. We provide *non-asymptotic* performance guarantees for the ℓ_1-regularized ML estimator in the Poisson regression model; that is, we not only prove the variable selection consistency and estimation consistency when the sample size increases to infinity, but also provide explicit bounds on estimation error and probability of correct variable selection for *any finite value* of the sample size.

2. We characterize the scaling of (n, p, s) that ensures variable selection consistency and estimation consistency, where n denotes the sample size, p denotes the parameter dimension, and s denotes the number of non-zero entries in the parameter. Our result shows that the ℓ_1-regularized ML estimator is consistent even when the parameter dimension increases exponentially with the sample size (cf. Corollary 4.1).

3. We derive novel inequalities for *self-concordant like* functions. Our framework enables a structured derivation of the consistency results, in the sense that most parts of our proof can be directly extended to statistical models involving a self-concordant like function, such as the logistic regression model [2]. We also develop computationally efficient algorithms to approximate the ℓ_1-regularized ML estimator for transmission tomography with the theory of self-concordant like functions; due to the page limit, we are not able to show the optimization theoretic results in this paper.

This work was supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof, SNF 200021-132548, SNF 200021-146750 and SNF CRSII2-147633.

Notations

We only point out some notations that might cause confusions without explicit definitions; other notations should be standard. For a vector $v \in \mathbb{R}^p$, we define the support function $\mathrm{supp}(v) := \{i : v_i \neq 0\}$. Let $\mathcal{S}$ be a subset of $\{1, \ldots, p\}$. We denote by $v_{\mathcal{S}}$ and $v_{\mathcal{S}^c}$ the sub-vector of v with entries indexed by $\mathcal{S}$ and $\{1, \ldots, p\} \setminus \mathcal{S}$, respectively. Similarly, for a matrix $A \in \mathbb{R}^{p \times p}$, we define the sub-matrices $A_{\mathcal{S},\mathcal{S}}$, $A_{\mathcal{S}^c,\mathcal{S}}$, etc. We write $\mathbb{R}^{\mathcal{S}}$ for the $|\mathcal{S}|$-dimensional subspace of $\mathbb{R}^p$ indexed by $\mathcal{S}$. We shall consider ℓ_p-norms $\|v\|_p := (\sum_{i=1}^p |v_i|^p)^{1/p}$, and $\|A\|_p$ denotes the operator norm of A induced by the ℓ_p-norm.

2. RELATED WORKS

The proof of our main theorem starts with a generalization of Wainwright's primal-dual witness approach [28] for the Gaussian linear regression model. However, due to the non-linearity of the Poisson regression model under consideration, we have to bound a residual term which will be defined clearly in Section 5, as other generalizations of Wainwright's work did [23, 24]. Unlike [23, 24], where specific techniques are developed for different statistical models, our proof is

structured, in the sense that it applies not only to the Poisson model, but can be easily adapted for any statistical model with a self-concordant like negative log-likelihood function.

The notion of self-concordant like functions was first proposed by Bach in [2], but the inequalities for self-concordant like functions presented in Section 6 are new. Our definition is slightly more general than the one in [2], as our definition is valid for any *norm* on $\mathbb{R}^p$ (cf. Definition 6.1), while the definition in [2] considers only the Euclidean norm. This allows us to obtain an estimation error bound in the ℓ_∞-norm, while the error bound in [2] is in the ℓ_2-norm. Note that a self-concordant like function is not necessarily self-concordant (cf. [19, 20] for the definition of self-concordant functions).

There are some general theoretical frameworks for proving the consistency of ℓ_1-regularized ML estimators, but they are not directly applicable for our purpose. The framework in [18] relies on the notion of *restricted strong convexity (RSC) condition*. However, the RSC condition alone does not guarantee variable selection consistency. A general framework for verifying variable selection consistency in generalized linear models is provided in [8]. The assumptions therein, however, are difficult to check for specific statistical models. We note that the only seemingly intractable assumptions in our paper are the irrepresentabillity condition and Assumption 1. The irrepresentability condition [28, 30], which has been proved to be almost necessary for the Gaussian linear regression model [28], and is also present in [8]. We shall discuss the validity of Assumption 1 in Section 4.

There are some other papers about regularized Poisson regression [13, 17, 22], but the regression model they considered is different from the one in this paper. Specifically, the negative log-likelihood functions in [13, 17, 22] are self-concordant but not self-concordant like; thus our analysis does not directly apply to their setting, and vice versa. Moreover, the variable selection consistency is not considered in [13, 17, 22].

We note that the optimal choice of the regularization coefficient based on given samples is in general an open problem [4]. Though in this paper we provide an explicit value of the regularization coefficient that ensures consistency, it is impractical since it requires at least the information of the true sparsity level (cf. (1)). In practice, one may apply the covariance penalty approach or cross validation procedures to estimate the optimal value of the regularization coefficient [1, 7].

3. PROBLEM FORMULATION

We formulate the Poisson regression model and the associated ℓ_1-regularized ML estimator in this section. Since the Poisson regression model finds applications in a variety of research areas, we deliberately keep the formulation general and ignore some physical constraints such as the positivity of the parameters in transmission tomography problems [10].

Let $\theta^\natural \in \mathbb{R}^p$ be the parameter to be estimated. Let $a_1, \dots, a_n$ be given vectors in $\mathbb{R}^p$. The measurement outcomes are samples $y_1, \dots, y_n$ of independent real-valued Poisson random variables $Y_1, \dots, Y_n$, with probability distribution given by

$$\mathbb{P}\{Y_i = y\} = \frac{\exp(-\lambda_i)\lambda_i^y}{y!}, \quad \text{for all } y \in \mathbb{N} \cup \{0\},$$

where $\lambda_i := \exp\left(-\left\langle a_i, \theta^\natural \right\rangle\right)$. Our aim is to estimate $\theta^\natural$ given $a_1, \dots, a_n$ and $y_1, \dots, y_n$.

The ℓ_1-regularized ML estimator, denoted by $\hat{\theta}_n$, is defined as

$$\hat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^p} \{f_n(\theta) + \rho_n g(\theta)\}$$

for some positive regularization coefficient ρ_n, where f_n is the normalized negative log-likelihood (up to some constant shift),

$$f_n(\theta) := \frac{1}{n}\sum_{i=1}^n \left[y_i \left\langle a_i, \theta \right\rangle + \exp\left(-\left\langle a_i, \theta \right\rangle\right)\right],$$

and g is the ℓ_1-norm, $g(\theta) := \|\theta\|_1$. The estimator $\hat{\theta}_n$ exists because $f_n + \rho_n g$ is coercive; it might not be unique, so we define it via an inclusion relation.

4. MAIN RESULT

In this section we discuss the assumptions, and then show the main theorem.

Assumption 1. The restricted Hessian $\left[\nabla^2 f_n(\theta^\natural)\right]_{\mathcal{S},\mathcal{S}}$ satisfies

$$\left\langle v, \left[\nabla^2 f_n(\theta^\natural)\right]_{\mathcal{S},\mathcal{S}} v \right\rangle \geq \mu \|v\|_1 \|v\|_\infty$$

with some $\mu > 0$, for any $v \in \mathbb{R}^s$.

Remark. Note this assumption also implies the positive definiteness of the restricted Hessian. By (3), f restricted on $\mathbb{R}^{\mathcal{S}}$ is strictly convex.

Assumption 1 is similar to the restricted eigenvalue condition [3], the compatibility condition [27], and the RIP-1 condition [12]. While Assumption 1 cannot be implied by any of them, numerical experiments strongly suggest that Assumption 1 might hold with high probability when $a_1, \dots, a_n$ are independent vectors of i.i.d. standard Gaussian random variables and $n \sim s\log(p/s)$. We leave the verification of Assumption 1 as a future work.

The second assumption is known as the irrepresentability condition [28, 30]. This condition has been proved to be "almost necessary" for the Gaussian linear regression model in [28].

Assumption 2. The Hessian satisfies

$$\left\| \left[\nabla^2 f_n(\theta^\natural)\right]_{\mathcal{S}^c,\mathcal{S}} \left[\nabla^2 f_n(\theta^\natural)\right]_{\mathcal{S},\mathcal{S}}^{-1} \right\|_\infty < 1 - \alpha$$

for some constant $\alpha \in (0, 1)$.

The final assumption is standard [2, 8, 28].

Assumption 3. Let $A \in \mathbb{R}^{n\times p}$ be the matrix whose i-th row is given by a_i. We assume that each column of A has ℓ_2-norm less than $\sqrt{n}$.

Theorem 4.1. *Suppose Assumptions 1–3 are satisfied. If ρ is chosen such that*

$$\rho_n \le \min\left\{\frac{\alpha\mu^2}{2\lambda_{\max}(4+\alpha)^2 s\left\|A_{\mathcal{S}}\right\|_\infty}, \frac{2\mu}{(4+\alpha)\left\|A_{\mathcal{S}}\right\|_\infty}\right\}. \tag{1}$$

then with probability at least $1-2p\exp\left[-\frac{\alpha\sqrt{n}\rho}{16\bar{\lambda}}\right]$, $\hat{\theta}$ satisfies $\hat{\theta}_{\mathcal{S}^c}=0$, and

$$\left\|\hat{\theta}_n-\theta^\natural\right\|_\infty \le \varepsilon_n := \left(\frac{4+\alpha}{2\mu}\right)\rho.$$

In addition, if $\theta_{\min} := \min_{i\in\mathcal{S}}\left\{\left|\theta_i^\natural\right|\right\} > \varepsilon_n$, then $\hat{\theta}_n$ recovers the sign pattern of $\theta^\natural$.

Consider the high-dimensional setting, where s and p can scale with n [8, 9, 28, 30].

Definition 4.1. A sequence of estimators $\{\hat{\theta}\}_{n\in\mathbb{N}}$ is consistent in variable selection if there exists $\{\rho_n\}_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty}\mathbb{P}\left\{\operatorname{supp}\left(\hat{\theta}_n\right)\neq\operatorname{supp}\left(\theta^\natural\right)\right\}=0.$$

Definition 4.2. A sequence of estimators $\{\hat{\theta}\}_{n\in\mathbb{N}}$ is consistent in estimation if there exists $\{\rho_n\}_{n\in\mathbb{N}}$ such that for any $\epsilon>0$,

$$\lim_{n\to\infty}\mathbb{P}\left\{\left\|\hat{\theta}_n-\theta^\natural\right\|>\epsilon\right\}=0,$$

where $\|\cdot\|$ is some norm on $\mathbb{R}^p$.

Choose ρ_n such that

$$n^{-1/2}\log p \ll \rho_n \ll s^{-1}\left\|A_{\mathcal{S}}\right\|_\infty^{-1}.$$

We obtain the following lemma.

Corollary 4.1. *If $\|A_{\mathcal{S}}\|_\infty s\log p \ll \sqrt{n}$ and $\theta_{\min} > \varepsilon_n$, then $\{\hat{\theta}\}_{n\in\mathbb{N}}$ is consistent in variable selection. If in addition, $\rho_n\to 0$ as $n\to\infty$, then $\{\hat{\theta}\}_{n\in\mathbb{N}}$ is also consistent in estimation.*

Note that μ and α are assumed to be constants.

5. SKETCH OF THE PROOF

Due to the page limit, we do not show the complete proof but briefly summarize the logical structure here. We omit the subscripts n in this section, and define $\mathcal{S} := \operatorname{supp}(\theta^\natural)$, to simplify the notation.

It is desirable to have $\hat{\theta}$ behave like the oracle estimator $\check{\theta}$, defined as

$$\check{\theta} := \arg\min_{\theta\in\mathbb{R}^p:\theta_{\mathcal{S}^c}=0}\left\{f(\theta)+\rho g(\theta)\right\}.$$

Since $f+\rho g$ restricted to $\mathbb{R}^{\mathcal{S}}$ is still coercive, $\check{\theta}$ exists. It can be verified that f is self-concordant like (cf. Definition 6.1). By Assumption 1 and (3), $f+\rho g$ is strictly convex on $\mathbb{R}^{\mathcal{S}}$ and thus $\check{\theta}$ is uniquely defined.

The following result is obtained by the primal-dual witness approach and can be proved as in [28], so we omit the proof here.

Lemma 5.1. *We have $\hat{\theta}=\check{\theta}$ if $\left\|\left[\nabla f(\check{\theta})\right]_{\mathcal{S}^c}\right\|_\infty<\rho$.*

By a Taylor series expansion on the first-order optimality condition of $\check{\theta}_n$ and the triangle inequality, the condition in Lemma 5.1 is satisfied if the irrepresentability condition (Assumption 2) holds, and [14]

$$\max\left\{\left\|\nabla f(\theta^\natural)\right\|_\infty, \|r\|_\infty\right\}\le\frac{\alpha}{4}\rho,$$

where the residual term r satisfies [29],

$$\|r\|_\infty \le \left\|\check{\theta}-\theta^\natural\right\|_\infty \sup_{t\in[0,1]}\left\{\left\|\nabla^2 f(\theta_t)-\nabla^2 f(\theta^\natural)\right\|_\infty\right\},$$

where $\theta_t := \theta^\natural + t(\check{\theta}-\theta^\natural)$. For the Gaussian linear regression model considered in [28], $r\equiv 0$, while the rest of our proof is mainly devoted to evaluating the non-zero residual term.

The norm of the residual term r is small if $\check{\theta}$ is close to $\theta^\natural$. This intuition is quantified by the following lemma.

Lemma 5.2. *We have $\|r\|_\infty\le(\alpha/4)\rho$ if*

$$\left\|\check{\theta}-\theta^\natural\right\|_\infty \le \min\left\{\sqrt{\frac{\alpha\rho}{8s\lambda_{\max}\left\|A_{\mathcal{S}}\right\|_\infty}}, \frac{1}{\left\|A_{\mathcal{S}}\right\|_\infty}\right\}. \tag{2}$$

where $\lambda_{\max} := \max\{\lambda_1,\ldots,\lambda_n\}$, and $A_{\mathcal{S}}$ denotes the submatrix of A whose columns are indexed by $\mathcal{S}$.

To verify (2) is equivalent to evaluating the estimation error of the oracle estimator $\check{\theta}$. In fact $\left\|\check{\theta}-\theta^\natural\right\|_\infty$ corresponds to the estimation error of the ℓ_1-regularized ML estimator under the classical $n>p$ setting, taking $\theta^\natural_{\mathcal{S}}$ as the parameter to be estimated.

Theorem 5.3. *We have*

$$\left\|\check{\theta}-\theta^\natural\right\|_\infty \le \frac{2}{\mu}\left(\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty+\rho\right),$$

given that

$$\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty+\rho\le\frac{\mu}{2\left\|A_{\mathcal{S}}\right\|_\infty}.$$

A short sketch of the proof of Theorem 5.3 is given in Section 6, where we make use of the fact that $\overline{f} := \mathbb{E}[f]$ is self-concordant like.

Bounding $\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty$ by $\left\|\nabla f(\theta^\natural)\right\|_\infty$, we conclude that $\hat{\theta} = \check{\theta}$ if the irrepresentability condition (Assumption 2) holds, $\left\|\nabla f(\theta^\natural)\right\|_\infty \leq (\alpha/4)\rho$, and

$$\rho \leq \min\left\{\frac{\alpha\mu^2}{2\lambda_{\max}(4+\alpha)^2 s\left\|A_{\mathcal{S}}\right\|_\infty}, \frac{2\mu}{(4+\alpha)\left\|A_{\mathcal{S}}\right\|_\infty}\right\}.$$

By applying Bernstein's inequality [15] to each element of $\nabla f(\theta^\natural)$ and the union bound, we show that $\left\|\nabla f(\theta^\natural)\right\|_\infty$ indeed concentrates around zero.

Lemma 5.4. *For any $t > 0$ and any $n > t^{-2}$,*

$$\mathbb{P}\left\{\left\|\nabla f_n(\theta^\natural)\right\|_\infty \geq t\right\} \leq 2p\exp\left[-\frac{\sqrt{n}t}{4\overline{\lambda}}\right].$$

where $\overline{\lambda}^2 := \max\left\{1, \lambda_{\max}^2\right\}$.

Remark. Note that the concentration behavior is not sub-Gaussian like, which is assumed in [8].

Theorem 4.1 and Corollary 4.1 follow by combining the intermediate results above.

6. TECHNICAL SUPPLEMENTS

6.1. Self-concordant like functions

Definition 6.1. A function $f : \mathrm{dom}(f) \subseteq \mathbb{R}^p \to \mathbb{R}$ is self-concordant like with parameter $M \geq 0$ with respect to a norm $\|\cdot\|$ on $\mathbb{R}^p$ if $\mathrm{dom}(f)$ is open, $f \in C^3(\mathrm{dom}(f))$, and

$$\left|D^3 f(x)[u,v,v]\right| \leq M\|u\| D^2 f(x)[v,v]$$

for any $x \in \mathrm{dom}(f)$ and $u, v \in \mathbb{R}^p$.

Remark. The special case where $\|\cdot\|$ is the ℓ_2-norm is considered in [2, 26].

Theorem 6.1. *Let $f : \mathrm{dom}(f) \subseteq \mathbb{R}^p \to \mathbb{R}$ be a self-concordant like function with parameter $M \geq 0$ with respect to a norm $\|\cdot\|$ on $\mathbb{R}^p$, and $x, y \in \mathrm{dom}(f)$. Define $r := M\|y - x\|$ and the local norm*

$$\|y - x\|_x := (D^2 f(x)[y-x, y-x])^{1/2}.$$

1. *Bounds on the Hessian:*

$$\exp(-r)\nabla^2 f(y) \leq \nabla^2 f(x) \leq \exp(r)\nabla^2 f(y). \quad (3)$$

2. *Bounds on the function value:*

$$\omega_*(r)\|y-x\|_x^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \omega(r)\|y-x\|_x^2, \quad (4)$$

where $\omega_*(r) := [\exp(-r) + r - 1]/r^2$ and $\omega(r) := [\exp(r) - r - 1]/r^2$.

Proof. Consider the function $\psi_u(t) := D^2 f(y_t)[u,u]$ for any $u \in \mathbb{R}^p$, where $y_t := x + t(y - x)$. Then we have

$$|\psi_u'(t)| = \left|D^3 f(y_t)[y-x, u, u]\right| \leq M\|y-x\|\psi_u(t),$$

and thus

$$\left|(\ln\psi_u(t))'\right| \leq M\|y-x\|.$$

We obtain (3) by integrating $(\ln\psi_u(t))'$ over $[0,1]$. We obtain (4) following the proof of Theorem 4.1.7 in [19]. □

6.2. Sketch of the Proof of Theorem 5.3

Define $\overline{f}(\theta) := \mathbb{E}[f(\theta)]$. Then we have $\overline{f}(\theta) - f(\theta) = \langle \nabla f(\theta^\natural), \theta\rangle$, and $\theta^\natural$ minimizes $\overline{f}(\theta)$.

By definition

$$f(\hat{\theta}) + \rho g(\hat{\theta}) \leq f(\theta^\natural) + \rho g(\theta^\natural),$$

which implies, by the triangle inequality,

$$\overline{f}(\hat{\theta}) - \overline{f}(\theta^\natural) \leq \left(\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty + \rho\right)\left\|\hat{\theta} - \theta^\natural\right\|_1. \quad (5)$$

It can be verified that for any $x, u, v \in \mathbb{R}^p$ satisfying $x_{\mathcal{S}^c} = u_{\mathcal{S}^c} = v_{\mathcal{S}^c} \equiv 0$,

$$\begin{aligned}\left|D^3\overline{f}(x)[u,v,v]\right| &\leq \max\left\{|(a_i)_{\mathcal{S}}, u|\right\} D^2\overline{f}(x)[v,v] \\ &\leq \|A_{\mathcal{S}}\|_\infty \|u\|_\infty D^2\overline{f}(x)[v,v],\end{aligned}$$

and thus $\overline{f}$, being restricted on $\mathbb{R}^{\mathcal{S}}$, is self-concordant like with parameter $\|A_{\mathcal{S}}\|_\infty$ with respect to the ℓ_∞-norm. By (4) and Assumption 1,

$$\overline{f}(\check{\theta}) - \overline{f}(\theta^\natural) \geq \frac{\mu}{M^2}\frac{\left\|\check{\theta} - \theta^\natural\right\|_1}{\left\|\check{\theta} - \theta^\natural\right\|_\infty}\left[\exp(-r) + r - 1\right], \quad (6)$$

where $r := M\left\|\check{\theta} - \theta^\natural\right\|_\infty$ and $M := \|A_{\mathcal{S}}\|_\infty$.

Combining (5) and (6), we obtain

$$\begin{aligned}&\exp\left(-M\left\|\check{\theta} - \theta^\natural\right\|_\infty\right) + M\left\|\check{\theta} - \theta^\natural\right\|_\infty - 1 \\ &\leq \frac{M}{\mu}\left(\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty + \rho\right) M\left\|\check{\theta} - \theta^\natural\right\|_\infty. \quad (7)\end{aligned}$$

Solving (7) directly, we obtain

$$M\left\|\check{\theta} - \theta^\natural\right\|_\infty \leq W_0\left[-\frac{1}{1-a}\exp\left(-\frac{1}{1-a}\right)\right],$$

where W_0 denotes the principal branch of the Lambert W function [6], and

$$a := \frac{M}{\mu}\left(\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty + \rho\right).$$

We simplify the solution by Theorem 3.2 in [25], and the theorem follows.

A. NUMERICAL EVIDENCE OF ASSUMPTION 1

Recall the definition of the matrix A in Assumption 3. By a direct calculation, we obtain

$$\nabla^2 f_n(\theta^\natural) = \frac{1}{n} A^T D A,$$

where $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $D_{i,i} = \operatorname{var} Y_i$ for all i. We assume that $D_{i,i} \geq \nu$ for some constant $\nu > 0$. Then

$$\left\langle v, \left[\nabla^2 f_n(\theta^\natural)\right]_{\mathcal{S},\mathcal{S}} v \right\rangle \geq \nu \left\langle v, A^T A v \right\rangle,$$

and we only need to verify whether there exists some $\tilde{\mu} > 0$ such that

$$\left\langle v, n^{-1} A_{\mathcal{S}}^T A_{\mathcal{S}} v \right\rangle \geq \tilde{\mu} \|v\|_1 \|v\|_\infty .$$

We would like to compare two values:

$$\mu_2^\natural := \min_{v\neq 0} \left\{ \frac{\left\langle v, n^{-1} A_{\mathcal{S}}^T A_{\mathcal{S}} v \right\rangle}{\|v\|_2^2} \right\},$$

$$\mu_{1,\infty}^\natural := \min_{v\neq 0} \left\{ \frac{\left\langle v, n^{-1} A_{\mathcal{S}}^T A_{\mathcal{S}} v \right\rangle}{\|v\|_1 \|v\|_\infty} \right\}.$$

The value of $\mu_2^\natural$ is the smallest restricted eigenvalue of $n^{-1}A^T A$, which will be close to 1 with high probability when n is large enough [11]; if $\mu_{1,\infty}^\natural$ is bounded away from 0, Assumption 1 holds.

We now describe the setup of the numerical experiment. Fix s, p, and n. First, we generate a random matrix $A \in \mathbb{R}^{n\times p}$ of independent and identically distributed standard Gaussian random variables. We generate 1000 independent samples $v_1, \ldots, v_{1000} \in \mathbb{R}^p$ of a random vector v. Each entry of v follows the uniform distribution on the interval $[-1, 1]$, and is statistically independent of other entries. Define

$$\mu_2 := \min_{1\leq i\leq 1000} \left\{ \frac{\left\langle v_i, n^{-1} A_{\mathcal{S}}^T A_{\mathcal{S}} v_i \right\rangle}{\|v_i\|_2^2} \right\},$$

$$\mu_{1,\infty} := \min_{1\leq i\leq 1000} \left\{ \frac{\left\langle v_i, n^{-1} A_{\mathcal{S}}^T A_{\mathcal{S}} v_i \right\rangle}{\|v_i\|_1 \|v_i\|_\infty} \right\}.$$

We approximate the values of $\mu_2^\natural$ and $\mu_{1,\infty}^\natural$ by μ_2 and $\mu_{1,\infty}$, respectively.

Figure 1 shows the values of μ_2 and $\mu_{1,\infty}$ with respect to varying s, the sparsity level, where we set $p = 10s$ and $n = 2s\log(p/s)$. We observe that $\mu_{1,\infty}$ is larger than 0.4 for all s (and thus the corresponding p and n) and follows the trend of μ_2. This provides a numerical evidence for the validity of Assumption 1.

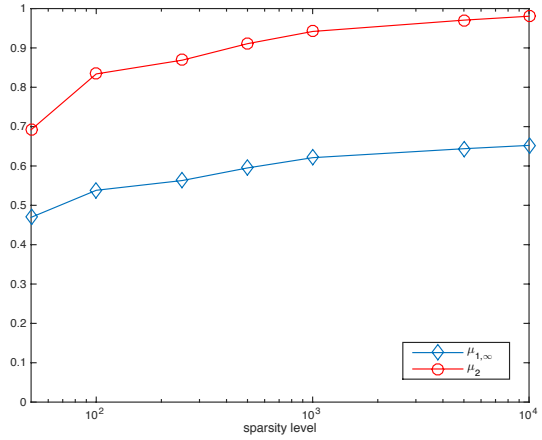

Fig. 1. Values of μ_2 and $\mu_{1,\infty}$ with respect to s, where we set $p = 10s$ and $n = 2s\log(p/s)$

B. PROOF OF MAIN RESULTS

Notation

Let $A \in \mathbb{R}^{p\times p}$ and $\mathcal{S}$ be a subset of $\{1, \ldots, p\}$. We denote by $A_{\mathcal{S}}$ the matrix in $\mathbb{R}^{n\times|\mathcal{S}|}$ whose columns are those in A indexed by $\mathcal{S}$. We omit the subscripts n for convenience; therefore, $f := f_n$, $\hat{\theta} := \hat{\theta}_n$, etc.

It is desirable to have $\hat{\theta}$ behave like the oracle estimator $\check{\theta}$, defined as

$$\check{\theta} := \arg \min_{\theta\in\mathbb{R}^p : \theta_{\mathcal{S}^c}=0} \{f(\theta) + \rho g(\theta)\}.$$

Since $f + \rho g$ restricted to $\mathbb{R}^{\mathcal{S}}$ is still coercive, $\check{\theta}$ exists.

Lemma B.1. *The estimator $\hat{\theta}$ is uniquely defined and equals $\check{\theta}$ if*

$$\left\| \left[\nabla f(\check{\theta})\right]_{\mathcal{S}^c} \right\|_\infty < \rho. \tag{8}$$

Proof. It can be verified that f, being restricted on $\mathbb{R}^{\mathcal{S}}$, is self-concordant like (cf. Definition 6.1). By Assumption 1 and (3), $f + \rho g$ is strictly convex on $\mathbb{R}^{\mathcal{S}}$ and thus $\check{\theta}$ is uniquely defined.

Since f being restricted on $\mathbb{R}^{\mathcal{S}}$ is strictly convex, $\check{\theta}$ is the only vector that satisfies

$$\left[\nabla f(\check{\theta})\right]_{\mathcal{S}} + \rho \check{z}_{\mathcal{S}} = 0 \tag{9}$$

with some $\check{z}_{\mathcal{S}}$ such that $\|\check{z}_{\mathcal{S}}\|_\infty \leq 1$. Suppose that (8) is satisfied. Then there exists $\check{z}_{\mathcal{S}^c}$ such that $\|\check{z}_{\mathcal{S}^c}\|_\infty < 1$ and

$$\nabla f(\check{\theta}) + \rho \check{z} = 0,$$

where $\check{z} := (\check{z}_{\mathcal{S}}, \check{z}_{\mathcal{S}^c})$. Therefore $\check{\theta}$ is a minimizer of $f + \rho g$.

Now we address the uniqueness of $\hat{\theta}$. By Lemma 1 (b) of [28] or Lemma 1 in [23], any minimizer $\tilde{\theta}$ of $f + \rho g$ also satisfies $\tilde{\theta}_{\mathcal{S}^c} = 0$. However, $\check{\theta}$ is the only vector v that minimizes $f(v) + \rho g(v)$ given the constraint $v_{\mathcal{S}^c} = 0$. Therefore $\hat{\theta} = \check{\theta}$ uniquely. □

By a Taylor series expansion on the first-order optimality condition of $\check{\theta}_n$ and the triangle inequality, inequality (8) is satisfied if the irrepresentability condition (2) holds, and [14]

$$\max\left\{ \left\|\nabla f(\theta^\natural)\right\|_\infty, \|r\|_\infty \right\} \leq \frac{\alpha}{4}\rho,$$

where the remainder term r satisfies [21]

$$\|r\|_\infty \le \sup_{t\in[0,1]} \left\{ \left\| \left[\nabla^2 f(\theta_t)\right]_{\mathcal{S}} - \left[\nabla^2 f(\theta^\natural)\right]_{\mathcal{S}} \right\|_\infty \right\} \left\| \check{\theta} - \theta^\natural \right\|_\infty , \tag{10}$$

where $\theta_t := \theta^\natural + t(\check{\theta} - \theta^\natural)$.

Remark. For the Gaussian linear regression model considered in [28], $r \equiv 0$, so the remaining derivation in [28] becomes much simpler.

By direct calculation we obtain

$$\nabla^2 f(\theta) = \frac{1}{n} A^T D(\theta) A,$$

where $A \in \mathbb{R}^{n\times p}$ denotes the matrix whose i-th row is a_i, and

$$D(\theta) := \operatorname{diag}(\exp(-\langle a_1, \theta\rangle), \ldots, \exp(-\langle a_n, \theta\rangle)) \in \mathbb{R}^{n\times n}.$$

Then

$$\begin{aligned} &\left\| \left[\nabla^2 f(\theta_t)\right]_{\mathcal{S}} - \left[\nabla^2 f(\theta^\natural)\right]_{\mathcal{S}} \right\|_\infty \\ &= \frac{1}{n} \left\| A^T \left[D(\theta_t) - D(\theta^\natural) \right] A_{\mathcal{S}} \right\|_\infty \\ &\le \frac{1}{n} \left\| A^T A_{\mathcal{S}} \right\|_\infty \max_{\substack{i\in\{1,\ldots,n\} \\ t\in[0,1]}} \left\{ \left| [D(\theta_t)]_{i,i} - \left[D(\theta^\natural)\right]_{i,i} \right| \right\}. \end{aligned}$$

Let ϕ_i denote the i-th column of A. By definition and Assumption 3,

$$\begin{aligned} \frac{1}{n} \left\| A^T A_{\mathcal{S}} \right\|_\infty &:= \frac{1}{n} \max_i \left\{ \sum_{j\in\mathcal{S}} |\phi_i, \phi_j| \right\} \\ &\le \frac{1}{n} \max_i \left\{ \sum_{j\in\mathcal{S}} \|\phi_i\|_2 \|\phi_j\|_2 \right\} \\ &\le \frac{1}{n} \max_i \left\{ \sum_{j\in\mathcal{S}} \sqrt{n}\sqrt{n} \right\} = s. \end{aligned}$$

Proposition B.2. *Suppose that*

$$\left\| \check{\theta} - \theta^\natural \right\|_\infty \le \|A_{\mathcal{S}}\|_\infty^{-1}. \tag{11}$$

Then

$$\begin{aligned} &\max_{\substack{i\in\{1,\ldots,n\} \\ t\in[0,1]}} \left\{ \left| [D(\theta_t)]_{i,i} - \left[D(\theta^\natural)\right]_{i,i} \right| \right\} \\ &\le 2\,\lambda_{\max} \|A_{\mathcal{S}}\|_\infty \left\| \check{\theta} - \theta^\natural \right\|_\infty , \end{aligned}$$

where $\lambda_{\max}$ denotes the maximum among $\lambda_1, \ldots, \lambda_n$.

Proof. By definition,

$$\begin{aligned} &\max_{\substack{i\in\{1,\ldots,n\} \\ t\in[0,1]}} \left\{ \left| [D(\theta_t)]_{i,i} - \left[D(\theta^\natural)\right]_{i,i} \right| \right\} \\ &= \max_{\substack{i\in\{1,\ldots,n\} \\ t\in[0,1]}} \left\{ \left| \exp(-\langle a_i, \theta_t\rangle) - \exp(-\langle a_i, \theta^\natural\rangle) \right| \right\} \\ &= \max_{\substack{i\in\{1,\ldots,n\} \\ t\in[0,1]}} \left\{ \lambda_i \left| 1 - \exp(-\langle a_i, \theta_t - \theta^\natural\rangle) \right| \right\}. \end{aligned}$$

Note that $|1 - \exp(-x)| \le 2|x|$ for any x such that $|x| \le 1$. Suppose

$$\left| \langle a_i, \theta_t - \theta^\natural \rangle \right| \le 1 \tag{12}$$

Then we have

$$\begin{aligned} \left| 1 - \exp(-\langle a_i, \theta_t - \theta^\natural\rangle) \right| &\le 2 \left| \langle a_i, \theta^\natural - \theta_t \rangle \right| \\ &\le 2 \left\| (a_i)_{\mathcal{S}} \right\|_1 \left\| \theta^\natural - \check{\theta} \right\|_\infty . \end{aligned}$$

Note that in the last inequality we write $(a_i)_{\mathcal{S}}$ instead of a_i because both θ_t and $\theta^\natural$ are in $\mathbb{R}^{\mathcal{S}}$. The condition (12) can be guaranteed by (11) by Hölder's inequality. Taking maximum with respect to i and t, the proposition follows. □

Then we have (cf. (10))

$$\|r\|_\infty \le 2s\lambda_{\max} \|A_{\mathcal{S}}\|_\infty \left\| \check{\theta} - \theta^\natural \right\|_\infty^2 ,$$

given that (11) is satisfied. This implies $\|r\|_\infty \le (\alpha/4)\rho$ if

$$\left\| \check{\theta} - \theta^\natural \right\|_\infty \le \min \left\{ \sqrt{\frac{\alpha\rho}{8s\lambda_{\max} \|A_{\mathcal{S}}\|_\infty}}, \frac{1}{\|A_{\mathcal{S}}\|_\infty} \right\}.$$

Now we evaluate $\left\| \check{\theta} - \theta^\natural \right\|_\infty$. This quantity is in fact the estimation error of the ℓ_1-regularized ML estimator in the conventional $n > p$ setting, taking $\theta^\natural_{\mathcal{S}}$ as the parameter to be estimated.

By properties of self-concordant like functions, we obtain the following theorem. The proof is given in Section C.

Theorem B.3. *We have*

$$\left\| \check{\theta} - \theta^\natural \right\|_\infty \le \frac{2}{\mu} \left(\left\| \left[\nabla f(\theta^\natural)\right]_{\mathcal{S}} \right\|_\infty + \rho \right),$$

given that

$$\left\| \left[\nabla f(\theta^\natural)\right]_{\mathcal{S}} \right\|_\infty + \rho \le \frac{\mu}{2\|A_{\mathcal{S}}\|_\infty}.$$

Therefore, $\|r\|_\infty \le (\alpha/4)\rho$ if

$$\begin{aligned} &\frac{2}{\mu} \left(\left\| \left[\nabla f(\theta^\natural)\right]_{\mathcal{S}} \right\|_\infty + \rho \right) \\ &\le \min \left\{ \sqrt{\frac{\alpha\rho}{8s\lambda_{\max} \|A_{\mathcal{S}}\|_\infty}}, \frac{1}{\|A_{\mathcal{S}}\|_\infty} \right\}, \end{aligned}$$

and

$$\left\| \left[\nabla f(\theta^\natural)\right]_{\mathcal{S}} \right\| + \rho \le \frac{\mu}{2\|A_{\mathcal{S}}\|_\infty}.$$

Under the assumption that $\left\| \nabla f(\theta^\natural) \right\|_\infty \le (\alpha/4)\rho$ and by the fact that $\left\| \left[\nabla f(\theta^\natural)\right]_{\mathcal{S}} \right\|_\infty \le \left\| \nabla f(\theta^\natural) \right\|_\infty$, it suffices to require

$$\begin{aligned} &\frac{2\rho}{\mu} \left(1 + \frac{\alpha}{4} \right) \\ &\le \min \left\{ \sqrt{\frac{\alpha\rho}{8s\lambda_{\max} \|A_{\mathcal{S}}\|_\infty}}, \frac{1}{\|A_{\mathcal{S}}\|_\infty} \right\}, \end{aligned}$$

and

$$\rho \leq \left(\frac{4}{4+\alpha}\right)\left(\frac{\mu}{2\left\|A_{\mathcal{S}}\right\|_{\infty}}\right).$$

To summarize, we have $\hat{\theta} = \check{\theta}$ if the irrepresentability condition (Assumption 2) holds, $\left\|\nabla f(\theta^{\natural})\right\|_{\infty} \leq (\alpha/4)\rho$, and

$$\rho \leq \min\left\{\frac{\alpha\mu^2}{2\lambda_{\max}(4+\alpha)^2 s\left\|A_{\mathcal{S}}\right\|_{\infty}}, \frac{2\mu}{(4+\alpha)\left\|A_{\mathcal{S}}\right\|_{\infty}}\right\}.$$

The concentration behavior of $\nabla f(\theta^{\natural})$ is characterized by the following lemma, the proof of which is given in Section D.

Lemma B.4. *For any $t > 0$ and any $n > t^{-2}$,*

$$\mathbb{P}\left\{\left\|\nabla f_n(\theta^{\natural})\right\|_{\infty} \geq t\right\} \leq 2p\exp\left[-\frac{\sqrt{n}t}{4\overline{\lambda}}\right].$$

where $\overline{\lambda}^2 := \max\left\{1, \lambda_{\max}^2\right\}$.

The intermediate results above can be summarized as the following theorem.

Theorem B.5. *Suppose Assumptions 1–3 are satisfied. If ρ is chosen such that*

$$\rho \leq \min\left\{\frac{\alpha\mu^2}{2\lambda_{\max}(4+\alpha)^2 s\left\|A_{\mathcal{S}}\right\|_{\infty}}, \frac{2\mu}{(4+\alpha)\left\|A_{\mathcal{S}}\right\|_{\infty}}\right\}.$$

then with probability at least $2p\exp\left[-\frac{\alpha\sqrt{n}\rho}{16\overline{\lambda}}\right]$, $\hat{\theta}$ satisfies $\hat{\theta}_{\mathcal{S}^c} = 0$, and

$$\left\|\hat{\theta} - \theta^{\natural}\right\|_{\infty} \leq \varepsilon_n := \frac{2}{\mu}\left(1 + \frac{\alpha}{4}\right)\rho.$$

In addition, if $\theta_{\min} := \min_{i\in\mathcal{S}}\left\{\left|\theta_i^{\natural}\right|\right\} > \varepsilon_n$, then $\hat{\theta}$ recovers the sign pattern of $\theta^{\natural}$.

Consider the high-dimensional setting, where s and p can scale with n [8, 9, 28, 30].

Definition B.1. The estimator $\hat{\theta}_n$ is consistent in variable selection if there exists ρ_n such that

$$\lim_{n\to\infty}\mathbb{P}\left\{\operatorname{supp}\left(\hat{\theta}_n\right) \neq \operatorname{supp}\left(\theta^{\natural}\right)\right\} = 0. \tag{13}$$

Definition B.2. The estimator $\hat{\theta}$ is consistent in estimation if there exists ρ_n such that for any $\epsilon > 0$,

$$\lim_{n\to\infty}\mathbb{P}\left\{\left\|\hat{\theta}_n - \theta^{\natural}\right\| > \epsilon\right\} = 0,$$

where $\|\cdot\|$ is some norm on $\mathbb{R}^p$.

Corollary B.1. *If $\left\|A_{\mathcal{S}}\right\|_{\infty} s\log p \ll \sqrt{n}$ and $\theta_{\min} > \varepsilon_n$, then $\hat{\theta}$ is consistent in variable selection. If in addition, $s\left\|A_{\mathcal{S}}\right\|_{\infty} \to \infty$ as $n \to \infty$, then $\hat{\theta}$ is also consistent in estimation.*

Proof. Choose ρ such that

$$n^{-1/2}\log p \ll \rho \ll s^{-1}\left\|A_{\mathcal{S}}\right\|_{\infty}^{-1}.$$

□

C. PROOF OF THEOREM 5.3

Define

$$\overline{f}(\theta) := \mathbb{E}\left[f(\theta)\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\lambda_i\left\langle a_i, \theta\right\rangle + \exp\left(-\left\langle a_i, \theta\right\rangle\right)\right].$$

For any $v \in \mathbb{R}^s$ define the zero-padding function $Z(v) := \tilde{v}$, where $\tilde{v}_{\mathcal{S}} = v$ and $\tilde{v}_{\mathcal{S}^c} = 0$. Define functions $f_{\mathcal{S}} := f \circ Z$ and $\overline{f}_{\mathcal{S}} := \overline{f} \circ Z$.

Then we have, for any $\theta \in \mathbb{R}^p$ satisfying $\theta_{\mathcal{S}^c} = 0$,

$$\begin{aligned} f(\theta) - \overline{f}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i - \lambda_i\right)\left\langle\left(a_i\right)_{\mathcal{S}}, \theta_{\mathcal{S}}\right\rangle\right] \\ &= \left\langle\left[\nabla f(\theta^{\natural})\right]_{\mathcal{S}}, \theta_{\mathcal{S}}\right\rangle. \end{aligned} \tag{14}$$

By definition,

$$\left[f(\check{\theta}) + \rho g(\check{\theta})\right] - \left[f(\theta^{\natural}) + \rho g(\theta^{\natural})\right] \leq 0,$$

which implies, by (14),

$$\begin{aligned} \overline{f}(\check{\theta}) - \overline{f}(\theta^{\natural}) \leq &-\left\langle\left[\nabla f(\theta^{\natural})\right]_{\mathcal{S}}, \check{\theta}_{\mathcal{S}} - \theta_{\mathcal{S}}^{\natural}\right\rangle - \\ &\rho\left[g(\check{\theta}) - g(\theta^{\natural})\right]. \end{aligned}$$

By Hölder's inequality and the triangle inequality,

$$\overline{f}(\check{\theta}) - \overline{f}(\theta^{\natural}) \leq \left(\left\|\left[\nabla f(\theta^{\natural})\right]_{\mathcal{S}}\right\|_{\infty} + \rho\right)\left\|\check{\theta} - \theta^{\natural}\right\|_1.$$

Since $\theta^{\natural}$ minimizes $\overline{f}$, the left-had side is non-negative, and the inequality is not trivial. This is equivalent to saying

$$\overline{f}_{\mathcal{S}}(\check{\theta}_{\mathcal{S}}) - \overline{f}_{\mathcal{S}}(\theta_{\mathcal{S}}^{\natural}) \leq \left(\left\|\left[\nabla f(\theta^{\natural})\right]_{\mathcal{S}}\right\|_{\infty} + \rho\right)\left\|\check{\theta} - \theta^{\natural}\right\|_1. \tag{15}$$

Lemma C.1. *Define $M := \left\|A_{\mathcal{S}}\right\|_{\infty}$. Then we have*

$$\begin{aligned} &\overline{f}_{\mathcal{S}}(\check{\theta}_{\mathcal{S}}) - \overline{f}_{\mathcal{S}}(\theta_{\mathcal{S}}^{\natural}) \\ &\quad\geq \frac{\mu}{M^2}\frac{\left\|\check{\theta} - \theta\right\|_1}{\left\|\check{\theta} - \theta\right\|_{\infty}}\left[\exp(-r) + r - 1\right], \end{aligned}$$

where $r := M\left\|\check{\theta} - \theta^{\natural}\right\|_{\infty}$ as in (4).

Proof. By direct calculation, for all $\theta, u, v \in \mathbb{R}^{|\mathcal{S}|}$,

$$\begin{aligned} &\left|D^3\overline{f}_{\mathcal{S}}(\theta)[u,v,v]\right| \\ &\quad= \frac{1}{n}\left|\sum_{i=1}^{n}\left\langle\left(a_i\right)_{\mathcal{S}}, u\right\rangle\left\langle\left(a_i\right)_{\mathcal{S}}, v\right\rangle^2\exp\left(-\left\langle\left(a_i\right)_{\mathcal{S}}, \theta\right\rangle\right)\right| \\ &\quad\leq \frac{1}{n}\sum_{i=1}^{n}\left|\left\langle\left(a_i\right)_{\mathcal{S}}, u\right\rangle\right|\left\langle\left(a_i\right)_{\mathcal{S}}, v\right\rangle^2\exp\left(-\left\langle\left(a_i\right)_{\mathcal{S}}, \theta\right\rangle\right) \\ &\quad\leq \left(\max_i\left|\left\langle\left(a_i\right)_{\mathcal{S}}, u\right\rangle\right|\right)D^2\overline{f}_{\mathcal{S}}(\theta)[v,v] \\ &\quad\leq \left\|A_{\mathcal{S}}\right\|_{\infty}\left\|u\right\|_{\infty}D^2\overline{f}_{\mathcal{S}}(\theta)[v,v], \end{aligned}$$

where the last equality is by Hölder's inequality. Therefore, $\overline{f}_{\mathcal{S}}$ is self-concordant like with parameter $M = \|A_{\mathcal{S}}\|_\infty$ with respect to the ℓ_∞-norm. Note that $\nabla \overline{f}_{\mathcal{S}}(\theta^\natural_{\mathcal{S}}) = 0$. The lemma follows by (4) and Assumption 1. □

Combining(15) and Lemma C.1, we obtain

$$\begin{aligned}&\exp\left(-M\left\|\check{\theta}-\theta^\natural\right\|_\infty\right)+M\left\|\check{\theta}-\theta^\natural\right\|_\infty-1\\ &\leq\frac{M}{\mu}\left(\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty+\rho\right)M\left\|\check{\theta}-\theta^\natural\right\|_\infty. \qquad (16)\end{aligned}$$

Solving (16) directly, we obtain

$$M\left\|\check{\theta}-\theta^\natural\right\|_\infty\leq W_0\left[-\frac{1}{1-a}\exp\left(-\frac{1}{1-a}\right)\right],$$

where W_0 denotes the principal branch of the Lambert W function [6], and

$$a:=\frac{M}{\mu}\left(\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty+\rho\right).$$

By Theorem 3.2 in [25], we can simplify the error bound to

$$M\left\|\check{\theta}-\theta^\natural\right\|_\infty\leq(1-a)^{-1}-2\leq(1-a)^{-1}-1.$$

Note that $(1-x)^{-1}-1\leq 2x$ for $x\in[0,1/2]$. We obtain

$$M\left\|\check{\theta}-\theta^\natural\right\|_\infty\leq\frac{2M}{\mu}\left(\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty+\rho\right),$$

given $a<1/2$; that is,

$$\left\|\check{\theta}-\theta^\natural\right\|_\infty\leq\frac{2}{\mu}\left(\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty+\rho\right),$$

given

$$\left\|\left[\nabla f(\theta^\natural)\right]_{\mathcal{S}}\right\|_\infty+\rho\leq\frac{\mu}{2\left\|A_{\mathcal{S}}\right\|_\infty}.$$

D. PROOF OF LEMMA 5.4

By direct calculation,

$$\begin{aligned}\left[\nabla f_n(\theta^\natural)\right]_j&=\frac{1}{n}\sum_{i=1}^n\left(Y_i-\mathbb{E}\left[Y_i\right]\right)(a_i)_j\\ &=\sum_{i=1}^n\left(\zeta_{i,j}-\mathbb{E}\left[\zeta_{i,j}\right]\right),\quad j\in\{1,\ldots,p\},\end{aligned}$$

where $\zeta_{i,j}:=n^{-1}(a_i)_jY_i$. Each entry of $\nabla f_n(\theta^\natural)$ is a sum of independent zero-mean random variables.

We shall make use of the Bernstein inequality [15].

Theorem D.1 (Bernstein's Inequality)**.** *Let $X_1,\ldots,X_n$ be independent real random variables. Suppose that there exist $\nu>0$ and $c>0$ such that $\sum_{i=1}^n\mathbb{E}\left[X_i^2\right]\leq\nu$, and*

$$\sum_{i=1}^n\mathbb{E}\left[|X_i|^q\right]\leq\frac{q!}{2}\nu c^{q-2}$$

for all integers $q\geq 3$. Then

$$\mathbb{P}\left\{\left|\sum_{i=1}^n\left(X_i-\mathbb{E}\left[X_i\right]\right)\right|\geq t\right\}\leq 2\exp\left[-\frac{t^2}{2(\nu+ct)}\right].$$

We now verify the moment conditions for random variables $\zeta_{1,j},\ldots,\zeta_{n,j}$ for all j.

Define $\lambda_i:=\exp\left(-\left\langle a_i,\theta^\natural\right\rangle\right)$, and ϕ_j as the j-th column of A. By definition $\mathbb{E}\left[\zeta_{i,j}^2\right]=(a_i)_j^2(\lambda_i^2+\lambda_i)/n^2$. Then, by Assumption 3,

$$\begin{aligned}\sum_{i=1}^n\mathbb{E}\left[\zeta_{i,j}^2\right]&=\frac{1}{n^2}\sum_{i=1}^n(a_j)_j^2(\lambda_i^2+\lambda_i)\\ &\leq\frac{1}{n^2}\left(\lambda_{\max}^2+\lambda_{\max}\right)\left\|\phi_j\right\|_2^2\\ &\leq\frac{1}{n}\left(\lambda_{\max}^2+\lambda_{\max}\right),\end{aligned}$$

where $\lambda_{\max}:=\max_{i\in\{1,\ldots,n\}}\{\lambda_i\}$. Thus it suffices to choose

$$\nu:=\begin{cases}2n^{-1} & ,\lambda_{\max}\leq 1\\ 2n^{-1}\lambda_{\max}^2 & ,\lambda_{\max}>1\end{cases}$$

in Theorem D.1. Also by definition, we have

$$\mathbb{E}\left[|\zeta_{i,j}|^q\right]=\frac{|A_{i,j}|^q}{n^q}\sum_{k=1}^q\lambda_i^q\left\{{q\atop k}\right\},$$

where $\left\{{q\atop i}\right\}$ denotes the Stirling number of the second kind. Then we have, for all $q\geq 3$,

$$\begin{aligned}\sum_{i=1}^n\mathbb{E}\left[|\zeta_{i,j}|^q\right]&=\frac{1}{n^q}\sum_{i=1}^n|A_{i,j}|^q\sum_{k=1}^q\lambda_i^q\left\{{q\atop k}\right\}\\ &\leq\frac{1}{n^q}\lambda_{\max}^q\left\|\phi_j\right\|_q^q\sum_{k=1}^q\left\{{q\atop k}\right\}\\ &\leq\frac{1}{n^q}\lambda_{\max}^q\left\|\phi_j\right\|_2^q\sum_{k=1}^q\left\{{q\atop k}\right\}\\ &\leq\frac{1}{n^{q/2}}\lambda_{\max}^q\sum_{k=1}^q\left\{{q\atop k}\right\}.\end{aligned}$$

Proposition D.2. *Let q be a positive integer. Then*

$$\sum_{k=1}^q\left\{{q\atop k}\right\}\leq q!.$$

Proof. We prove by induction.

By definition we have, for positive integers q and k,

$$\left\{{q+1\atop k}\right\}=k\left\{{q\atop k}\right\}+\left\{{q\atop k-1}\right\}.$$

It is easy to verify that the proposition holds for $q=1$. Suppose that the proposition holds for $q=r\geq 1$. Then for

$q = r + 1$,

$$\sum_{k=1}^{r+1} \left\{ {r+1 \atop k} \right\} = \sum_{k=1}^{r} k \left\{ {r \atop k} \right\} + \sum_{k=1}^{r} \left\{ {r \atop k} \right\} \leq (r+1) \sum_{k=1}^{r} \left\{ {r \atop k} \right\} \leq (r+1)!.$$

□

For the case where $\lambda_{\max} \leq 1$, we have

$$\sum_{i=1}^{n} \mathbb{E}\left[|\zeta_{i,j}|^q\right] \leq \frac{1}{n^{q/2}} \lambda_{\max}^q \sum_{k=1}^{q} \left\{ {q \atop k} \right\} \leq \frac{q!}{2} \nu n^{(q-2)/2},$$

and thus it suffices to choose $c = n^{-1/2}$. For the case where $\lambda_{\max} > 1$, we have

$$\sum_{i=1}^{n} \mathbb{E}\left[\left|\zeta_{i,j}^q\right|\right] \leq \frac{1}{n^{q/2}} \lambda_{\max}^q \sum_{k=1}^{q} \left\{ {q \atop k} \right\} \leq \frac{q!}{2} \nu \lambda_{\max}^{q-2} n^{(q-2)/2},$$

and thus it suffices to choose $c = \lambda_{\max} n^{-1/2}$.

By the union bound and Bernstein's inequality, we obtain, for any $t > 0$,

$$\mathbb{P}\left\{ \left\| \nabla f_n(\theta^\natural) \right\|_\infty \geq t \right\} \leq 2p \exp\left[-\frac{t^2}{2\overline{\lambda}\left(n^{-1} + n^{-1/2} t\right)} \right],$$

where $\overline{\lambda}^2 := \max\left\{1, \lambda_{\max}^2\right\}$.

E. REFERENCES

[1] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Stat. Surv.*, vol. 4, pp. 40–79, 2010.

[2] F. Bach, "Self-concordant analysis for logistic regression," *Electron. J. Stat.*, vol. 4, pp. 384–414, 2010.

[3] P. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Stat.*, vol. 37, no. 4, pp. 1705–1732, 2009.

[4] P. J. Bickel, "Regularization in statistics," *Test*, vol. 15, no. 2, pp. 271–344, 2006.

[5] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*, 2nd ed. New York: Cambridge Univ. Press, 2013.

[6] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, pp. 329–359, 1996.

[7] B. Efron, "The estimation of prediction error: Covariance penalities and cross-validation," *J. Am. Stat. Assoc.*, vol. 99, no. 467, pp. 619–632, Sep. 2004.

[8] J. Fan and J. Lv, "Nonconcave penalized likelihood with NP-dimensionality," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5467–5484, Aug. 2011.

[9] J. Fan and H. Peng, "Nonconcave penalized likelihood with a diverging number of parameters," *Ann. Stat.*, vol. 32, no. 3, pp. 928–961, 2004.

[10] J. A. Fessler, "Statistical image reconstruction methods for transmission tomography," in *Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis*, M. Sonka and J. M. Fitzpatrick, Eds. Bellingham: SPIE Press, 2000, ch. 1, pp. 1–70.

[11] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Basel: Birkhäuser, 2013.

[12] A. Gilbert and P. Indyk, "Sparse recovery using sparse matrices," *Proc. IEEE*, vol. 98, no. 6, pp. 937–947, Jun. 2010.

[13] X. Jiang, G. Raskutti, and R. Willett, "Minimax optimal rates for Poisson inverse problems with physical constraints," 2014, arXiv:1403.6532v1 [math.ST].

[14] Y.-H. Li, J. Scarlett, P. Ravikumar, and V. Cevher, "Sparsistency of ℓ_1-regularized M-estimators," in *18th Inf. Conf. Artificial Intelligence and Statistics*, 2015.

[15] P. Massart, *Concentration Inequalities and Model Selection*. Berlin: Springer-Verl., 2007.

[16] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. London: Chapman and Hall, 1989.

[17] D. Motamedvaziri, M. H. Rohban, and V. Saligrama, "Sparse signal recovery under Poisson statistics," 2014, arXiv:1307.4666v2 [math.ST].

[18] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

[19] Y. Nesterov, *Introductory Lectures on Convex Optimization*. Boston, MA: Kluwer, 2004.

[20] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*. Philadelphia, PA: SIAM, 1994.

[21] B. T. Polyak, *Introduction to Optimization*. New York: Optimization Softw., Inc., 1987.

[22] M. Raginsky, R. M. Willett, Z. T. Harmany, and R. F. Marcia, "Compressed sensing performance bounds under Poisson noise," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 3990–2010, Aug. 2010.

[23] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Ann. Stat.*, vol. 38, no. 3, pp. 1287–1319, 2010.

[24] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Stat.*, vol. 5, pp. 935–980, 2011.

[25] S. M. Stewart, "On certain inequalities involving the Lambert W function," *J. Inequal. Pure Appl. Math.*, vol. 10, 2009.

[26] Q. Tran-Dinh, Y.-H. Li, and V. Cevher, "Composite convex minimization involving self-concordant-like cost functions," 2015, arXiv:1502.01068 [math.OC].

[27] S. van de Geer, "The deterministic Lasso," Seminar für Statistik, Eidgenössische Technische Hochschule, Research Report No. 140, 2007.

[28] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[29] E. Zeidler, *Applied Functional Analysis: Main Principles and Their Applications*. New York, NY: Springer-Verl., 1995.

[30] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, 2006.