Sparsistency of ℓ_1-Regularized M-Estimators

Yen-Huan Li
LIONS, EPFL

Jonathan Scarlett
LIONS, EPFL

Pradeep Ravikumar
University of Texas at Austin

Volkan Cevher
LIONS, EPFL

Abstract

We consider the model selection consistency or *sparsistency* of a broad set of ℓ_1-regularized M-estimators for linear and non-linear statistical models in a unified fashion. For this purpose, we propose the local structured smoothness condition (LSSC) on the loss function. We provide a general result giving deterministic sufficient conditions for sparsistency in terms of the regularization parameter, ambient dimension, sparsity level, and number of measurements. We show that several important statistical models have M-estimators that indeed satisfy the LSSC, and as a result, the sparsistency guarantees for the corresponding ℓ_1-regularized M-estimators can be derived as simple applications of our main theorem.

1 Introduction

This paper studies the class of ℓ_1-regularized M-estimators for *sparse* high-dimensional estimation [Bühlmann and van de Geer, 2011]. A key motivation for adopting such estimators is sparse model selection, that is, selecting the important subset of entries of a high-dimensional parameter based on random observations. We study the conditions for the reliable recovery of the sparsity pattern, commonly known as model selection consistency or *sparsistency*.

For the specific case of sparse linear regression, the ℓ_1-regularized least squares estimator has received considerable attention. With respect to sparsistency, results have been obtained for both the noiseless case (e.g., [Candes and Tao, 2005, Donoho, 2006, Donoho and Tanner, 2005]) and the noisy case [Meinshausen and Bühlmann, 2006, Wainwright, 2009, Zhao and Yu, 2006]. While sparsistency results have been obtained for ℓ_1-regularized M-estimators on some *specific* non-linear models such as logistic regression and Gaussian Markov random field models [Bach, 2010, Bunea, 2008, Lam and Fan, 2009, Meinshausen and Bühlmann, 2006, Ravikumar et al., 2010, Ravikumar et al., 2011], *general* techniques with broad applicability are largely lacking.

Performing a general sparsistency analysis requires the identification of general properties of statistical models, and their corresponding M-estimators, that can be exploited to obtain strong performance guarantees. In this paper, we introduce the *local structured smoothness condition* (LSSC) condition (Definition 3.1), which controls the smoothness of the objective function in a particular structured set. We illustrate how the LSSC enables us to address a broad set of sparsistency results in a unified fashion, including logistic regression, gamma regression, and graph selection. We explicitly check the LSSC for these statistical models, and as in previous works [Fan and Lv, 2011, Fan and Peng, 2004, Ravikumar et al., 2010, Ravikumar et al., 2011, Wainwright, 2009, Zhao and Yu, 2006], we derive sample complexity bounds for the high-dimensional setting, where the ambient dimension and sparsity level are allowed to scale with the number of samples.

To the best of our knowledge, the first work to study the sparsistency of a broad class of models was that of [Fan and Lv, 2011] for generalized linear models; however, the technical assumptions therein appear to be difficult to check for specific models, thus making their application difficult. Another related work is [Lee et al., 2014]; in Section 7, we compare the two, and discuss a key advantage of our approach.

The paper is organized as follows. We specify the problem setup in Section 2. We introduce the LSSC in Section 3, and give several examples of functions satisfying the LSSC in Section 4. In Section 5, we present the main theorem of this paper, namely, sufficient conditions for an ℓ_1-regularized M-estimator to successfully recover the support. Sparsistency results for four dif-

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38. Copyright 2015 by the authors.

ferent statistical models are established in Section 6 as corollaries of our main result. In Section 7, we present further discussions of our results, and list some directions for future research. The proofs of our results can be found in the supplementary material.

2 Problem Setup

We consider a general statistical modeling setting where we are given n independent samples $\{y_i\}_{i=1}^n$ drawn from some distribution $\mathbb{P}$ with a sparse parameter $\beta^* := \beta(\mathbb{P}) \in \mathbb{R}^p$ that has at most s non-zero entries. We are interested in estimating this sparse parameter β^* given the n samples via an ℓ_1-regularized M-estimator of the form

$$\hat{\beta}_n := \arg\min_{\beta \in \mathbb{R}^p} L_n(\beta) + \tau_n \left\| \beta \right\|_1 , \tag{1}$$

where L_n is some convex function, and $\tau_n > 0$ is a regularization parameter.

We mention here a special case of this model that has broad applications in machine learning. For fixed vectors $x_1, \ldots, x_n$ in $\mathbb{R}^p$, suppose that we are given realizations $y_1, \ldots, y_n$ of independent random variables $Y_1, \ldots, Y_n$ in $\mathbb{R}$. We assume that each Y_i follows a probability distribution P_{θ_i} parametrized only by θ_i, where $\theta_i := \langle x_i, \beta^* \rangle$ for some sparse parameter $\beta^* \in \mathbb{R}^p$. Then it is natural to consider the ℓ_1-regularized maximum-likelihood estimator

$$\hat{\beta}_n := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i; \beta, x_i) + \tau_n \left\| \beta \right\|_1 ,$$

where ℓ denotes the negative log-likelihood at y_i given x_i and β. Thus, we obtain (1) with $L_n(\beta) := \frac{1}{n} \sum_{i=1}^n \ell(y_i; \beta, x_i)$.

There are of course many other examples; to name one other, we mention the graphical learning problem, where we want to learn a sparse concentration matrix of a vector-valued random variable. In this setting, we also arrive at the formulation (1), where L_n is the negative log-likelihood of the data [Ravikumar et al., 2011].

We focus on the *sparsistency* of $\hat{\beta}_n$; roughly speaking, an estimator $\hat{\beta}_n$ is sparsistent if it recovers the support of β^* with high probability when the number of samples n is large enough.

Definition 2.1 (Sparsistency)**.** A sequence of estimators $\{\hat{\beta}_n\}_{n=1}^\infty$ is called *sparsistent* if

$$\lim_{n \to \infty} \mathsf{P} \left\{ \operatorname{supp} \hat{\beta}_n \neq \operatorname{supp} \beta^* \right\} = 0.$$

The main result of this paper is that, if the function L is convex and satisfies the LSSC, and certain assumptions analogous to those used for linear models (see [Wainwright, 2009]) hold true, then the ℓ_1-regularized M-estimator $\hat{\beta}_n$ in (1) is sparsistent under suitable conditions on the regularization parameter τ_n and the triplet (n, p, s). We allow for the case of diverging dimensions [Fan and Lv, 2011, Ravikumar et al., 2010, Ravikumar et al., 2011, Wainwright, 2009, Zhao and Yu, 2006], where p grows exponentially with n.

Notations and Basic Definitions

Fix $v \in \mathbb{R}^p$, and let $\mathcal{P} = \{1, \ldots, p\}$. For any $\mathcal{S} \subseteq \mathcal{P}$, the notation $v_{\mathcal{S}}$ denotes the sub-vector of v on $\mathcal{S}$, and the notation $v_{\mathcal{S}^c}$ denotes the sub-vector $v_{\mathcal{P} \setminus \mathcal{S}}$. For $i \in \mathcal{P}$, the notation v_i denotes $v_{\{i\}}$. We denote the support set of v by $\operatorname{supp} v$, defined as $\operatorname{supp} v = \{i : v_i \neq 0, i \in \mathcal{P}\}$. The notation $\operatorname{sign} v$ denotes the vector $(\operatorname{sign} v_1, \ldots, \operatorname{sign} v_p)$, where $\operatorname{sign} v_i = v_i \left| v_i \right|^{-1}$ if $v_i \neq 0$, and $\operatorname{sign} v_i = 0$ otherwise, for all $i \in \mathcal{P}$. We denote the transpose of v by v^T, and the ℓ_q-norm of v by $\left\| v \right\|_q$ for $q \in [1, +\infty]$. For $u, v \in \mathbb{R}^p$, the notation $\langle u, v \rangle$ denotes $\sum_{i=1}^p u_i v_i$.

For $A \in \mathbb{R}^{p \times p}$, the notations $A_{\mathcal{S},\mathcal{S}}$, $A_{\mathcal{S}^c,\mathcal{S}}$, $\operatorname{supp} A$, $\operatorname{sign} A$, and A^T are defined analogously to the vector case. The notation $\left\| A \right\|_q$ denotes the operator norm induced by the vector ℓ_q-norm; in particular, $\left\| A \right\|_2$ denotes the spectral norm of A.

Let X be a real-valued random variable. We denote the expectation and variance of X by $\mathsf{E}\, X$ and $\mathsf{var}\, X$, respectively. The probability of an event $\mathcal{E}$ is denoted by $\mathsf{P}\, \mathcal{E}$.

Let f be a vector-valued function with domain $\operatorname{dom} f \subseteq \mathbb{R}^p$. The notations ∇f and $\nabla^2 f$ denote the gradient and Hessian mapping of f, respectively. The notation $f \in \mathcal{C}^k(\operatorname{dom} f)$ means that f is k-times continuously differentiable on $\operatorname{dom} f$. For a given function $f \in \mathcal{C}^k(\operatorname{dom} f)$, its k-th order Fréchet derivative at $x \in \operatorname{dom} f$ is denoted by $D^k f(x)$, which is a multilinear symmetric form [Zeidler, 1995]. The following special cases summarize how to compute all of the quantities related to the Fréchet derivative in this paper:

1. The first order Fréchet derivative is simply the gradient mapping; therefore, $Df(x)[u] = \langle \nabla f(x), u \rangle$ for all $u \in \mathbb{R}^p$.

2. The second order Fréchet derivative is the Hessian mapping; therefore, $D^2 f(x)[u, v] = \langle u, \nabla^2 f(x) v \rangle$ for all $u, v \in \mathbb{R}^p$.

3. The third order Fréchet derivative is defined as follows. We first define the 2-linear form (matrix)

$D^3 f(x)[u] := \lim_{t \to 0} \frac{\nabla^2 f(x+tu) - \nabla^2 f(x)}{t}$. Then

$$\begin{aligned} D^3 f(x)[u,v,w] &= \left(D^3 f(x)[u] \right) [v,w] \\ &= \left\langle v, (D^3 f(x)[u]) w \right\rangle . \end{aligned}$$

We then define the 1-linear form (vector) $D^3 f(x)[u,v]$ to be the unique vector such that $\langle D^3 f(x)[u,v], w \rangle = D^3 f(x)[u,v,w]$ for all vectors w in $\mathbb{R}^p$.

4. When the arguments are the same, we simply have $D^k f(x)[u,\dots,u] = \left. \frac{d^k \phi_u(t)}{dt^k} \right|_{t=0}$, where $\phi_u(t) := f(x+tu)$.

3 Local Structured Smoothness Condition

The following definition provides the key property of convex functions that will be exploited in the subsequent sparsistency analysis.

Definition 3.1 (Local Structured Smoothness Condition (LSSC))**.** Consider a function $f \in C^3(\mathrm{dom\,f})$ with domain $\mathrm{dom}\, f \subseteq \mathbb{R}^p$. Fix $x^* \in \mathrm{dom}\, f$, and let $\mathcal{N}_{x^*}$ be an open set in $\mathrm{dom}\, f$ containing x^*. The function f satisfies the $(x^*, \mathcal{N}_{x^*})$-LSSC with parameter $K \geq 0$ if

$$\left\| D^3 f(x^* + \delta)[u,u] \right\|_\infty \leq K \left\| u \right\|_2^2 ,$$

for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$, and for all $u \in \mathbb{R}^p$ such that $u_{\mathcal{S}^c} = 0$, where $\mathcal{S} := \mathrm{supp}\, x^*$.

Note that $D^3 f(x^* + \delta)[u,u]$ is a 1-linear form, so $\| \cdot \|_\infty$ in Definition 3.1 is the vector ℓ_∞-norm. The following equivalent characterization follows immediately.

Proposition 3.1. *The function f satisfies the $(x^*, \mathcal{N}_{x^*})$-LSSC with parameter $K \geq 0$ if and only if*

$$\left| D^3 f(x^* + \delta)[u,u,e_j] \right| \leq K \left\| u \right\|_2^2 , \qquad (2)$$

for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$, for all $u \in \mathbb{R}^p$ such that $u_{\mathcal{S}^c} = 0$, where $\mathcal{S} := \mathrm{supp}\, x^*$, and for all $j \in \{1,\dots,p\}$, where e_j is the standard basis vector with 1 in the j-th position and 0s elsewhere.

As we will see in the next section, this equivalent characterization is useful when verifying the LSSC for a given M-estimator.

Since differentiation is a linear operator, the LSSC is preserved under linear combinations with positive coefficients, as is stated formally in the following lemma.

Lemma 3.2. *Let f_1 satisfy the $(x, \mathcal{N}_1)$-LSSC with parameter K_1, and f_2 satisfy the $(x, \mathcal{N}_2)$-LSSC with parameter K_2. Let α and β be two positive real numbers. The function $f := \alpha f_1 + \beta f_2$ satisfies the $(x, \mathcal{N}_x)$-LSSC with parameter K, where $\mathcal{N}_x := \mathcal{N}_1 \cap \mathcal{N}_2$, and $K := \alpha K_1 + \beta K_2$.*

We conclude this section by briefly discussing the connection of the LSSC with other conditions. The following result, Proposition 9.1.1 of [Nesterov and Nemirovskii, 1994], will be useful here and throughout the paper.

Proposition 3.3. *Let A be a 3-linear symmetric form on $(\mathbb{R}^p)^3$, and B be a positive-semidefinite 2-linear symmetric form on $(\mathbb{R}^p)^2$. If*

$$|A[u,u,u]| \leq B[u,u]^{3/2}$$

for all $u \in \mathbb{R}^p$, then

$$|A[u,v,w]| \leq B[u,u]^{1/2} B[v,v]^{1/2} B[w,w]^{1/2}$$

for all $u, v, w \in \mathbb{R}^p$.

This proposition shows that the condition in (2) *without structural constraints on u and e_j* is equivalent to the statement that

$$\left| D^3 f(x^* + \delta)[u,v,w] \right| \leq K \left\| u \right\|_2 \left\| v \right\|_2 \left\| w \right\|_2 \qquad (3)$$

for all $u, v, w \in \mathbb{R}^p$. In the supplementary material, we show that (3) holds for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$ if and only if

$$\left\| D^2 f(x^* + \delta) - D^2 f(x^*) \right\|_2 \leq K \left\| \delta \right\|_2 , \qquad (4)$$

for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$. The latter condition is simply the local Lipschitz continuity of the Hessian of f. This is why we consider our condition a *local structured smoothness* condition, with structural constraints on the inputs of the $D^3 f(x^* + \delta)$ operator.

The preceding observations reveal that (3), or the equivalent formulation (4), is more restrictive than the LSSC. That is, (3) implies the LSSC, while the reverse is not true in general.

4 Examples

In this section, we provide some examples of functions that satisfy the LSSC.

Example 4.1. Suppose that $f(\beta) := \| y - X\beta \|_2^2$ for some fixed $y \in \mathbb{R}^p$ and $X \in \mathbb{R}^{n \times p}$. Since $D^3 f(\beta) \equiv 0$ everywhere, the function f satisfies the $(\beta^*, \mathcal{N}_{\beta^*})$-LSSC with parameter $K = 0$ for any $\beta^* \in \mathbb{R}^p$ and any open set $\mathcal{N}_{\beta^*} \subseteq \mathbb{R}^p$ that contains β^*. This function appears in the negative-likelihood in the Gaussian regression model.

Example 4.2. Let $f(\beta) := \langle x, \beta \rangle - \ln \langle x, \beta \rangle$ for some fixed $x \in \mathbb{R}^p$. We show that, for any fixed $\beta^* \in \mathrm{dom}\, f$ such that $\beta^*_{\mathcal{S}^c} = 0$, there exists some non-negative K and some open set $\mathcal{N}_{\beta^*}$ such that f satisfies the $(\beta^*, \mathcal{N}_{\beta^*})$-LSSC with parameter K. This function appears in the negative log-likelihood in gamma regression with the canonical link function.

By a direct differentiation, we obtain for all $u \in \mathbb{R}^p$ that

$$\begin{aligned} &\left| D^3 f(\beta^* + \delta)[u,u,u] \right| \\ &\quad = 2 \left| 1 + \gamma \right|^{-3} \left\{ D^2 f(\beta^*)[u,u] \right\}^{3/2}, \end{aligned} \tag{5}$$

where

$$\gamma := \frac{\langle x, \delta \rangle}{\langle x, \beta^* \rangle},$$

Combining this with Proposition 3.3, we have for each standard basis vector e_j that

$$\begin{aligned} &\left| D^3 f(\beta^* + \delta)[u,u,e_j] \right| \\ &\quad \le 2 \left| 1 + \gamma \right|^{-3} D^2 f(\beta^*)[u,u] \left\{ D^2 f(\beta^*)[e_j, e_j] \right\}^{1/2} \\ &\quad \le 2 \left(1 - |\gamma| \right)^{-3} D^2 f(\beta^*)[u,u] \left\{ D^2 f(\beta^*)[e_j, e_j] \right\}^{1/2}, \end{aligned}$$

if $|\gamma| \le 1$. Now define $\mathcal{S} := \operatorname{supp} \beta^*$, and suppose that $u_{\mathcal{S}^c} = \delta_{\mathcal{S}^c} = 0$, and that

$$\|\delta\|_2 \le \frac{\langle x, \beta^* \rangle}{(1 + \kappa) \|x_{\mathcal{S}}\|_2}$$

for some $\kappa > 0$. By the Cauchy-Schwartz inequality, it immediately follows that $|\gamma| \le (1+\kappa)^{-1} < 1$, and thus $\beta^* + \delta$ is in $\operatorname{dom} f$. Moreover, using this bound on $|\gamma|$, we can further upper bound $|D^3 f|$ as

$$\left| D^3 f(\beta^* + \delta)[u,u,e_j] \right| \le 2 \left(1 + \kappa^{-1} \right)^3 \lambda_{\max} d_{\max}^{1/2} \|u\|_2^2,$$

where $\lambda_{\max}$ is the maximum restricted eigenvalue of $D^2 f(\beta^*)$ defined as

$$\lambda_{\max} := \sup_{\substack{\|u\|_2 \le 1 \\ u_{\mathcal{S}^c} = 0}} D^2 f(\beta^*)[u,u],$$

and $d_{\max}$ denotes the maximum diagonal entry of $\nabla^2 f(\beta^*)$. Therefore, f satisfies the $(\beta^*, \mathcal{N}_{\beta^*})$-LSSC with parameter $K := 2(1+\kappa^{-1})^3 \lambda_{\max} d_{\max}^{1/2}$, where

$$\mathcal{N}_{\beta^*} := \left\{ \beta^* + \delta : \|\delta\|_2 \le \frac{\langle x, \beta^* \rangle}{(1+\kappa)\|x_{\mathcal{S}}\|_2}, \delta \in \mathbb{R}^p \right\}.$$

Example 4.3. Consider the function $f(\Theta) = \operatorname{Tr}(X\Theta) - \ln \det \Theta$ with a fixed $X \in \mathbb{R}^{p \times p}$, and with $\operatorname{dom} f := \{\Theta \in \mathbb{R}^{p \times p} : \Theta > 0\}$. We show that, for any fixed $\Theta^* \in \operatorname{dom} f$, there exists some non-negative K and some open set $\mathcal{N}_{\Theta^*}$ such that f satisfies the $(\Theta^*, \mathcal{N}_{\Theta^*})$-LSSC with parameter K. This function appears as the negative log-likelihood in the Gaussian graphical learning problem.

Note that the previous definitions (in particular, Definition 3.1), should be interpreted here as being taken with respect to the vectorizations of the relevant matrices.

It is already known that f is standard self-concordant [Nesterov, 2004]; that is,

$$\left| D^3 f(\Theta^* + \Delta)[U,U,U] \right| \le 2 \left\{ D^2 f(\Theta^* + \Delta)[U,U] \right\}^{3/2},$$

for all $U \in \mathbb{R}^{p \times p}$ and all $\Delta \in \mathbb{R}^{p \times p}$ such that $\Theta^* + \Delta \in \operatorname{dom} f$. This implies, by Proposition 3.3,

$$\begin{aligned} \left| D^3 f(\Theta^* + \Delta)[U,U,V] \right| &\le 2 \left\{ D^2 f(\Theta^* + \Delta)[U,U] \right\} \\ &\qquad \left\{ D^2 f(\Theta^* + \Delta)[V,V] \right\}^{1/2}, \end{aligned}$$

for all $U, V \in \mathbb{R}^{p \times p}$, and all $\Delta \in \mathbb{R}^{p \times p}$ such that $\Theta^* + \Delta \in \operatorname{dom} f$.

Moreover, by a direct differentiation,

$$\begin{aligned} \left\| D^2 f(\Theta^* + \Delta) \right\|_2 &= \left\| (\Theta^* + \Delta)^{-1} \otimes (\Theta^* + \Delta)^{-1} \right\|_2 \\ &= \left\| (\Theta^* + \Delta)^{-1} \right\|_2^2. \end{aligned}$$

Fix a positive constant κ, and suppose that we choose Δ such that $\|\Delta\|_F \le (1+\kappa)^{-1} \rho_{\min}$, where $\rho_{\min}$ denotes the smallest eigenvalue of Θ^*. Since $\|\Delta\|_2 \le \|\Delta\|_F$, it follows that $\|\Delta\|_2 \le (1+\kappa)^{-1} \rho_{\min}$, and, by Weyl's theorem [Horn and Johnson, 1985],

$$\left\| (\Theta^* + \Delta)^{-1} \right\|_2 \ge \frac{\kappa}{1+\kappa} \rho_{\min}.$$

Combining the preceding observations, it follows that f satisfies the $(\Theta^*, \mathcal{N}_{\Theta^*})$-LSSC with parameter $K := 2\kappa^{-3}(1+\kappa)^3 \rho_{\min}^{-3}$, where

$$\begin{aligned} \mathcal{N}_{\Theta^*} = \Big\{ \Theta^* + \Delta : \|\Delta\|_F < \frac{1}{1+\kappa} \rho_{\min}, \\ \Delta = \Delta^T, \Delta \in \mathbb{R}^{p \times p} \Big\}. \end{aligned}$$

Here we have not exploited the special structure of U in Definition 3.1 (namely, $u_{\mathcal{S}^c} = 0$), though conceivably the constant K could improve by doing so. Note that $\mathcal{N}_{\Theta^*} \subset \operatorname{dom} f$ and $\mathcal{N}_{\Theta^*}$ is convex.

5 Deterministic Sufficient Conditions

We are now in a position to state the main result of this paper, whose proof can be found in the supplementary material.

Let $\beta^* \in \mathbb{R}^p$ be the true parameter, and let $\mathcal{S} = \{i : (\beta^*)_i \neq 0\}$ be its support set. Define the "genie-aided" estimator with exact support information:

$$\check{\beta}_n \in \arg \min_{\beta \in \mathbb{R}^p : \beta_{\mathcal{S}^c} = 0} L_n(\beta) + \tau_n \|\beta\|_1. \tag{6}$$

Theorem 5.1. *Suppose that $\check{\beta}_n$ is uniquely defined. Then the ℓ_1-regularized estimator $\hat{\beta}_n$ defined in (1)* *uniquely exists, successfully recovers the sign pattern, i.e.,* $\operatorname{sign}\hat{\beta}_n = \operatorname{sign}\beta^*$*, and satisfies the error bound*

$$\left\|\hat{\beta}_n - \beta^*\right\|_2 \leq r_n := \frac{\alpha+4}{\lambda_{\min}}\sqrt{s}\tau_n, \tag{7}$$

if the following conditions hold true.

1. (Local structured smoothness condition) L_n *is convex, three times continuously differentiable, and satisfies the* $(\beta^*, \mathcal{N}_{\beta^*})$*-LSSC with parameter* $K \geq 0$*, for some convex* $\mathcal{N}_{\beta^*} \subseteq \operatorname{dom} L_n$.
2. (Positive definite restricted Hessian) *The restricted Hessian at* β^* *satisfies* $\left[\nabla^2 L_n(\beta^*)\right]_{\mathcal{S},\mathcal{S}} \geq \lambda_{\min} I$ *for some* $\lambda_{\min} > 0$.
3. (Irrepresentablility condition) *For some* $\alpha \in (0,1]$*, it holds that*
$$\left\|\left[\nabla^2 L_n(\beta^*)\right]_{\mathcal{S}^c,\mathcal{S}} \left[\nabla^2 L_n(\beta^*)\right]_{\mathcal{S},\mathcal{S}}^{-1}\right\|_\infty < 1-\alpha. \tag{8}$$
4. (Beta-min condition) *The smallest non-zero entry of* β *satisfies*
$$\beta_{\min} := \min\left\{|(\beta^*)_k| : k \in \mathcal{S}\right\} > r_n, \tag{9}$$
where r_n *is defined in (7).*
5. *The regularization parameter* τ_n *satisfies*
$$\tau_n < \frac{\lambda_{\min}^2}{4(\alpha+4)^2}\frac{\alpha}{Ks}. \tag{10}$$
6. *The gradient of* L_n *at* β^* *satisfies*
$$\|\nabla L_n(\beta^*)\|_\infty \leq \frac{\alpha}{4}\tau_n. \tag{11}$$
7. *The relation* $\mathcal{B}_{r_n} \subseteq \mathcal{N}_{\beta^*}$ *holds, where*
$$\mathcal{B}_{r_n} := \left\{\beta \in \mathbb{R}^p : \|\beta_n - \beta^*\|_2 \leq r_n, \beta_{\mathcal{S}^c} = 0\right\}$$
and r_n *is defined in (7).*

As mentioned previously, the first condition is the key assumption permitting us to perform a general analysis. The second, third, and forth assumptions are analogous to those appearing in the literature for sparse linear regression. We refer to [Bühlmann and van de Geer, 2011] for a systematic discussion of these conditions.[1]

The remaining conditions determine the interplay between τ_n , n, p, and s. Whether the relation $\mathcal{B}_{r_n} \subseteq \mathcal{N}_{\beta^*}$ holds depends on the specific $\mathcal{N}_{\beta^*}$ that one can derive for the given loss function L_n. Whether the upper bound on $\|\nabla L_n(\beta^*)\|_\infty$ holds depends on the concentration of measure behavior of $\nabla L_n(\beta^*)$, which usually concentrates around 0. In the next section, we will give concrete examples for the high-dimensional setting, where p and s scale with n.

Of course, $\operatorname{sign}\hat{\beta}_n = \operatorname{sign}\beta^*$ implies that $\operatorname{supp}\hat{\beta}_n = \operatorname{supp}\beta^*$, i.e. successful support recovery.

6 Applications

In this section, we provide several applications of Theorem 5.1, presenting concrete bounds on the sample complexity in each case. We defer the full proofs of the results in this section to the supplementary material. However, in each case, we present here the most important step of the proof, namely, verifying the LSSC.

Note that instead of the classical setting where only the sample size n increases, we consider the high-dimensional setting, where the ambient dimension p and the sparsity level s are allowed to grow with n [Fan and Lv, 2011, Fan and Peng, 2004, Ravikumar et al., 2010, Ravikumar et al., 2011, Wainwright, 2009, Zhao and Yu, 2006].

6.1 Linear Regression

We first consider the linear regression model with additive sub-Gaussian noise. This setting trivially fits into our theoretical framework.

Definition 6.1 (Sub-Gaussian Random Variables)**.** A zero-mean real-valued random variable Z is *sub-Gaussian* with parameter $c > 0$ if

$$\mathsf{E}\exp(tZ) \leq \exp\left(\frac{c^2t^2}{2}\right)$$

for all $t \in \mathbb{R}$.

Let $\mathcal{X}_n := \{x_1, \ldots, x_n\} \subset \mathbb{R}^n$ be given. Define the matrix $X_n \in \mathbb{R}^{n\times p}$ such that the i-th row of X_n is x_i. We assume that the elements in $\mathcal{X}_n$ are normalized such that each column of X has ℓ_2-norm less than or equal to $\sqrt{n}$. Let $W_1, \ldots, W_n$ be independent sub-Gaussian random variables with parameter c, and define $Y_i := \langle x_i, \beta^*\rangle + W_i$.

We consider the ℓ_1-regularized M-estimator of the form (1), with

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\left(Y_i - \langle x_i, \beta\rangle\right)^2.$$

As shown in the first example of Section 4, L_n satisfies the LSSC with parameter $K = 0$ everywhere in

[1] Equation (8) is sometimes called the *incoherence condition* [Wainwright, 2009].

$\mathbb{R}^p$. Therefore, the condition on τ_n in (10) is trivially satisfied, as is the final condition listed in the theorem.

By a direct calculation, we have

$$\nabla L_n(\beta^*) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathsf{E}\, Y_i)x_i.$$

By the union bound and the standard concentration inequality for sub-Gaussian random variables [Boucheron et al., 2013],

$$\begin{aligned}&\mathsf{P}\left\{\|\nabla L_n(\beta^*)\|_\infty \ge \frac{\alpha\tau_n}{4}\right\}\\ &\quad\le \sum_{i=1}^{p}\mathsf{P}\left\{\left|[\nabla L_n(\beta^*)]_i\right| \ge \frac{\alpha\tau_n}{4}\right\}\\ &\quad\le 2p\exp\left(-cnt^2\right)\Big|_{t=\frac{\alpha\tau_n}{4}}.\end{aligned}$$

Since $[D^2 L_n(\beta)]_{\mathcal{S},\mathcal{S}} = [D^2 L_n(\beta^*)]_{\mathcal{S},\mathcal{S}}$ is positive definite for all $\beta \in \mathbb{R}^p$ by the second assumption of Theorem 5.1, $\check{\beta}_n$ uniquely exists, and Theorem 5.1 is applicable. By choosing τ_n sufficiently large that the above bound decays to zero, we obtain the following.

Corollary 6.1. *For the linear regression problem described above, suppose that assumptions 2 to 4 of Theorem 5.1 hold for some $\lambda_{\min}$ and α bounded away from zero.*[2] *If $s\log p \ll n$, and we choose $\tau_n \gg (n^{-1}\log p)^{1/2}$, then the ℓ_1-regularized maximum likelihood estimator is sparsistent.*

Observe that this recovers the scaling law given in [Wainwright, 2009] for the linear regression model.

6.2 Logistic Regression

Let $\mathcal{X}_n := \{x_1, \ldots, x_n\} \subset \mathbb{R}^n$ be given. As in Section 6.1, we assume that $\sum_{j=1}^{n}(x_i)_j^2 \le n$ for all $i \in \{1, \ldots, p\}$.

Let $\beta^* \in \mathbb{R}^p$ be sparse, and define $\mathcal{S} := \operatorname{supp}\beta^*$. We are interested in estimating β^* given $\mathcal{X}_n$ and $\mathcal{Y}_n := \{y_1, \ldots, y_n\}$, where each y_i is the realization of a Bernoulli random variable Y_i with

$$\mathsf{P}\left\{Y_i = 1\right\} = 1 - \mathsf{P}\left\{Y_i = 0\right\} = \frac{1}{1+\exp\left(-\langle x_i, \beta^*\rangle\right)}.$$

The random variables $Y_1, \ldots, Y_n$ are assumed to be independent.

We consider the ℓ_1-regularized maximum-likelihood estimator of the form (1) with

$$L_n(\beta) := \frac{1}{n}\sum_{i=1}^{n}\ln\left\{1+\exp\left[-(2Y_i-1)\langle x_i, \beta\rangle\right]\right\}.$$

Define $\ell_i(\beta) = \ln\left[1+\exp\left(-(2y_i-1)\langle x_i, \beta\rangle\right)\right]$. The cases $y_i = 0$ and $y_i = 1$ are handled similarly, so we focus on the latter. A direct differentiation yields the following (this is most easily verified for $u = v$):

$$\begin{aligned}&\left|D^3\ell_i(\beta^*+\delta)[u,u,v]\right|\\ &\quad= \frac{\left|1-\exp\left(-\langle x_i, \beta^*+\delta\rangle\right)\right|}{1+\exp\left(-\langle x_i, \beta^*+\delta\rangle\right)}\left|\langle x_i, v\rangle\right| D^2\ell_i(\beta^*+\delta)[u,u]\\ &\quad\le \left|\langle x_i, v\rangle\right| D^2\ell_i(\beta^*+\delta)[u,u],\end{aligned}$$

and

$$\begin{aligned}D^2\ell_i(\beta)[u,u] &= \frac{\exp\left(-\langle x_i, \beta\rangle\right)\langle x_i, u\rangle^2}{\left[1+\exp\left(-\langle x_i, \beta\rangle\right)\right]^2}\\ &\le \frac{1}{4}\langle x_i, u\rangle^2\end{aligned}$$

for all $\beta \in \mathbb{R}^p$. The last inequality follows since the function $\frac{z}{(1+z)^2}$ has a maximum value of $\frac{1}{4}$ for $z \ge 0$. It follows that

$$\begin{aligned}\left|D^3\ell_i(\beta^*+\delta)[u,u,v]\right| &\le \frac{1}{4}\left|\langle x_i, v\rangle\right|\left|\langle x_i, u\rangle\right|^2\\ &\le \frac{1}{4}\left\|(x_i)_{\mathcal{S}}\right\|_2^2\left\|x_i\right\|_\infty\left\|u\right\|_2^3,\end{aligned}$$

for any $u \in \mathbb{R}^p$ such that $u_{\mathcal{S}^c} = 0$, and for any v equal to some standard basis vector e_j. Hence, L_n satisfies the $(\beta^*, \mathcal{N}_{\beta^*})$-LSSC with parameter $K = (1/4)\nu_n^2\gamma_n$, where

$$\begin{aligned}\nu_n &:= \max_i \left\|(x_i)_{\mathcal{S}}\right\|_2,\\ \gamma_n &:= \max_i \left\|x_i\right\|_\infty,\end{aligned}$$

and where $\mathcal{N}_{\beta^*}$ can be any fixed open convex neighborhood of β^* in $\mathbb{R}^p$.

Corollary 6.2. *For the logistic regression problem described above, suppose that assumptions 2 to 4 of Theorem 5.1 hold for some $\lambda_{\min}$ and α bounded away from zero. If we choose $\tau_n \gg (n^{-1}\log p)^{1/2}$, and s and p such that $s^2\left(\log p\right)\nu_n^4\gamma_n^2 \ll n$, then the ℓ_1-regularized maximum-likelihood estimator is sparsistent.*

In [Bunea, 2008], a scaling law of the form $s \ll \frac{\sqrt{n}}{(\log n)^2}$ is given, but the result is restricted to the case that p grows polynomially with n. The result in [Bach, 2010] yields the scaling $s^2(\log p)\overline{\nu}_n^{-2} \ll n$, where $\overline{\nu}_n := \max\left\{\left\|x_i\right\|_2\right\}$. It should be noted that $\overline{\nu}_n$ is generally significantly larger than ν_n and γ_n; for example, for i.i.d. Gaussian vectors, these scale on average as $O(\sqrt{p})$, $O(\sqrt{s})$ and $O(1)$, respectively. Our result recovers the same dependence of n on s and p as that in [Bach, 2010], but removes the dependence on $\overline{\nu}_n$. Of course, we do not restrict p to grow polynomially with n.

[2] For all of the examples in this section, these assumptions are independent of the data, and we can thus talk about them being satisfied *deterministically*.

6.3 Gamma Regression

Let $\mathcal{X}_n := \{x_1, \dots, x_n\} \subset \mathbb{R}^n$ be given. We again assume that $\sum_{j=1}^n (x_i)_j^2 \leq n$ for all $i \in \{1, \dots, p\}$.

Let $\beta^* \in \mathbb{R}^p$ be sparse, and define $\mathcal{S} := \operatorname{supp} \beta^*$. We are interested in estimating β^* given $\mathcal{X}_n$ and $\mathcal{Y}_n := \{y_1, \dots, y_n\}$, where each y_i is the realization of a gamma random variable Y_i with known shape parameter $k > 0$ and unknown scale parameter $\theta_i = k^{-1} \langle x_i, \beta^* \rangle^{-1}$. The corresponding density function is of the form $\frac{1}{\Gamma(k)\theta_i^k} y_i^{k-1} e^{-\frac{y_i}{\theta_i}}$.

We assume that

$$\langle x_i, \beta^* \rangle \geq \mu_n \quad \forall i \in \{1, \dots, n\} \tag{12}$$

for some $\mu_n > 0$, so θ_i is always well-defined. Moreover, the random variables $Y_1, \dots, Y_n$ are assumed to be independent.

We consider the ℓ_1-regularized maximum-likelihood estimator of the form (1) with

$$L_n(\beta) := \frac{1}{n} \sum_{i=1}^n \left[-\ln \langle x_i, \beta \rangle + Y_i \langle x_i, \beta \rangle \right].$$

Note that θ_i only enters the log-likelihood via constant terms not containing β; these have been omitted, as they do not affect the estimation.

Defining $\ell_i(\beta) = -\ln \langle x_i, \beta \rangle + y_i \langle x_i, \beta \rangle$, we obtain the following for all $u \in \mathbb{R}^p$ such that $u_{\mathcal{S}^c} = 0$, using the Cauchy-Schwartz inequality and (12):

$$\begin{aligned} D^2 \ell_i(\beta^*)[u, u] = \frac{\langle x_i, u \rangle^2}{\langle x_i, \beta^* \rangle^2} &\leq \frac{\|(x_i)_{\mathcal{S}}\|_2^2}{\langle x_i, \beta^* \rangle^2} \|u\|_2^2 \\ &\leq \frac{1}{\mu_n^2} \|u\|_2^2 \|(x_i)_{\mathcal{S}}\|_2^2 . \end{aligned}$$

Thus, the largest restricted eigenvalue of $D^2 \ell_i(\beta^*)$ is upper bounded by $\mu_n^{-2} \nu_n^2$, where $\nu_n = \max_i \{\|(x_i)_{\mathcal{S}}\|_2\}$. Similarly, we obtain

$$D^2 \ell_i(\beta^*)[e_j, e_j] \leq \frac{1}{\mu_n^2} \|x_i\|_\infty^2 ,$$

for any standard basis vector e_j. Thus, the largest diagonal entry of $D^2 \ell_i(\beta^*)$ is upper bounded by $\mu_n^{-2} \gamma_n^2$, where $\gamma_n = \max_i \|x_i\|_\infty$.

Fix $\kappa > 0$. By Example 4.2 and Lemma 3.2, L_n satisfies the $(\beta^*, \mathcal{N}_{\beta^*})$-LSSC with parameter $K = 2(1 + \kappa^{-1})^3 \mu_n^{-3} \nu_n^2 \gamma_n$, and

$$\mathcal{N}_{\beta^*} = \left\{ \beta^* + \delta : \|\delta\|_2 < \frac{\mu_n}{(1 + \kappa)\nu_n}, \delta \in \mathbb{R}^p \right\}.$$

Corollary 6.3. *Consider the gamma regression problem as described above, and suppose that assumptions 2 to 4 of Theorem 5.1 hold for some $\lambda_{\min}$, and α bounded away from zero. If $\tau_n \gg \sqrt{n}^{-1} \log p$ and $s^2 (\log p)^2 \mu_n^{-6} \nu_n^4 \gamma_n^2 \ll n$, then the ℓ_1-regularized maximum likelihood estimator is sparsistent.*

To the best of our knowledge, this is the first sparsistency result for gamma regression.

6.4 Graphical Model Learning

Let $\Theta^* \in \mathbb{R}^{p \times p}$ be a positive-definite matrix. We assume there are at most s non-zero entries in Θ^*, and let $\mathcal{S}$ denote its support set. Let $X_1, \dots, X_n$ be independent p-dimensional random vectors generated according to a common distribution with mean zero and covariance matrix $\Sigma^* := (\Theta^*)^{-1}$. We are interested in recovering the support of Θ^* given $X_1, \dots, X_n$.

We assume that each $(\Sigma_{i,i})^{-1/2} X_{i,i}$ is sub-Gaussian with parameter $c > 0$, and that $\Sigma_{i,i}$ is bounded above by a constant κ_{Σ^*}, for all $i \in \{1, \dots, p\}$. Let $\rho_{\min}$ denote the smallest eigenvalue of Θ^*.

We consider the ℓ_1-regularized M-estimator of the form (1), given by

$$\hat{\Theta}_n := \arg \min_{\Theta} \left\{ L_n(\Theta) + \tau_n |\Theta|_1 : \Theta > 0, \Theta \in \mathbb{R}^{p \times p} \right\}.$$

Here $|\Theta|_1$ denotes the entry-wise ℓ_1-norm, i.e., $|\Theta|_1 = \sum_{(i,j) \in \{1,\dots,p\}^2} |\Theta_{i,j}|$ and

$$L_n(\Theta) = \operatorname{Tr} \left(\hat{\Sigma}_n \Theta \right) - \log \det \Theta,$$

where $\hat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^T$ is the sample covariance matrix.

Fix $\kappa > 0$. By Example 4.3, we know that L_n satisfies the $(\Theta^*, \mathcal{N}_{\Theta^*})$-LSSC with parameter $2\kappa^{-3}(1+\kappa)^3 \rho_{\min}^{-3}$, where

$$\begin{aligned} \mathcal{N}_{\Theta^*} := \Big\{ \Theta^* + \Delta : \|\Delta\|_F < \frac{1}{1 + \kappa} \rho_{\min}, \\ \Delta = \Delta^T, \Delta \in \mathbb{R}^{p \times p} \Big\}, \end{aligned}$$

where $\rho_{\min}$ denotes the smallest eigenvalue of Θ^*.

The beta-min condition can be written as

$$\min \left\{ \Theta_{i,j}^* : \Theta_{i,j}^* \neq 0, (i,j) \in \{1, \dots, p\}^2 \right\} > r_n.$$

We now have the following.

Corollary 6.4. *Consider the graphical model selection problem described above, and suppose the above assumptions and assumptions 2 to 4 of Theorem 5.1 hold for some c, κ_{Σ^*}, $\rho_{\min}$, $\lambda_{\min}$, and α bounded away from zero. If $\tau_n \gg (n^{-1} \log p)^{1/2}$ and $s^2 \log p \ll n$, the ℓ_1-regularized M-estimator $\hat{\Theta}_n$ is sparsistent.*

Corollary 6.4 is for graphical learning on general sparse networks, as we only put a constraint on s. Several previous works have instead imposed structural constraints on the maximum degree of each node; e.g. see [Ravikumar et al., 2011]. Since this model requires additional structural assumptions beyond sparsity alone, it is outside the scope of our theoretical framework.

7 Discussion

Our work bears some resemblance to the independent work of [Lee et al., 2014]. The smoothness condition therein is in fact the *non-structured* condition in (4). From the discussion in Section 3, we see that our condition is less restrictive. As a consequence, both analyses lead to scaling laws of the form $n \gg K^2 s^2 (\log p)^\gamma$ for some $\gamma > 0$ for generalized linear models, but the corresponding definitions of K differ significantly. Eliminating the dependence of K on p requires additional non-trivial extensions of the framework in [Lee et al., 2014], whereas in our framework the desired independence is immediate (e.g. see the logistic and gamma regression examples).

The derivation of estimation error bounds such as (7) (as opposed to full sparsistency) usually only requires some kind of local *restricted strong convexity* (RSC) condition [Negahban et al., 2012] on L_n. It is interesting to note that in this paper, it suffices for sparsistency to assume only the LSSC and the positive definiteness of the restricted Hessian at the true parameter. It would be interesting to derive connections between the LSSC and such local RSC conditions, which in turn may shed light on whether the LSSC is necessary to derive sparsistency results, or whether a weaker condition may suffice.

The framework presented here considers general sparse parameters. It is of great theoretical and practical importance to sharpen this framework for structured sparse parameters, e.g., group sparsity, and graphical model learning for networks with bounded degrees.

Acknowledgements

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.

References

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

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford Univ. Press, Oxford.

[Bühlmann and van de Geer, 2011] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data.* Springer, Berlin.

[Bunea, 2008] Bunea, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.*, 2:1153–1194.

[Candes and Tao, 2005] Candes, E. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215.

[Donoho, 2006] Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306.

[Donoho and Tanner, 2005] Donoho, D. L. and Tanner, J. M. (2005). Neighborliness of randomly-projected simplices in high dimensions. *Proc. Nat. Acad. Sci.*, 102(27):9452–9457.

[Fan and Lv, 2011] Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inf. Theory*, 57(8):5467–5484.

[Fan and Peng, 2004] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.*, 32(3):928–961.

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

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

[Lee et al., 2014] Lee, J. D., Sun, Y., and Taylor, J. E. (2014). On model selection consistency of M-estimators with geometrically decomposable penalties. arXiv:1305.7477v7.

[Meinshausen and Bühlmann, 2006] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, 34:1436–1462.

[Negahban et al., 2012] 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. *Stat. Sci.*, 27(4):538–557.

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

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

[Ravikumar et al., 2010] Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Stat.*, 38(3):1287–1319.

[Ravikumar et al., 2011] Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980.

[Wainwright, 2009] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory*, 55(5):2183–2202.

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

[Zhao and Yu, 2006] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563.

Sparsistency of ℓ_1-Regularized M-Estimators: Supplementary Material

Yen-Huan Li
LIONS, EPFL

Jonathan Scarlett
LIONS, EPFL

Pradeep Ravikumar
University of Texas at Austin

Volkan Cevher
LIONS, EPFL

1 Auxiliary Result for the Non-Structured Case

In this section, we prove the following claim made in Section 3. Note that, in contrast to the main definition of the LSSC, the vectors here are *not* necessarily structured.

Proposition 1.1. *Consider a function $f \in \mathcal{C}^3(\mathrm{dom\,f})$ with domain $\mathrm{dom}\, f \subseteq \mathbb{R}^p$. Fix $x^* \in \mathrm{dom}\, f$, and let $\mathcal{N}_{x^*}$ be an open set in $\mathrm{dom}\, f$ containing x^*. Let $K \geq 0$. The following statements are equivalent.*

1. *$D^2 f(x)$ is locally Lipschitz continuous with respect to x^*; that is,*
$$\left\| D^2 f(x^* + \delta) - D^2 f(x^*) \right\|_2 \leq K \left\| \delta \right\|_2, \qquad (1)$$
for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$.

2. *$D^3 f(x)$ is locally bounded; that is,*
$$\left| D^3 f(x^* + \delta)[u, v, w] \right| \leq K \left\| u \right\|_2 \left\| v \right\|_2 \left\| w \right\|_2 \qquad (2)$$
for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$, and for all $u, v, w \in \mathbb{R}^p$.

Proof. Suppose that (1) holds. By Proposition 3.3, it suffices to prove that
$$\left| D^3 f(x^* + \delta)[u, u, u] \right| \leq K \left\| u \right\|_2^3$$
for all $u \in \mathbb{R}^p$. By definition, we have
$$\begin{aligned} \left| D^3 f(x^* + \delta)[u, u, u] \right| &= \left| \langle u, Hu \rangle \right| \\ &\leq \left\| H \right\|_2 \left\| u \right\|^2, \end{aligned}$$
where
$$H := \lim_{t \to 0} \frac{D^2 f(x^* + \delta + tu) - D^2 f(x^* + \delta)}{t}.$$
We therefore have (2) since $\left\| H \right\|_2 \leq K \left\| \delta \right\|_2$ by (1).

Conversely, suppose that (2) holds. We have the following Taylor expansion [Zeidler, 1995]:
$$D^2 f(x^* + \delta) = D^2 f(x^*) + \int_0^1 D^3 f(x_t)[\delta] \, dt,$$
where $x_t := x^* + t\delta$. We also have from (2) and the definition of the spectral norm that $\left\| D^3 f(x^* + \delta)[\delta] \right\|_2 \leq K \left\| u \right\|_2$, and hence
$$\begin{aligned} &\left\| D^2 f(x^* + \delta) - D^2 f(x^*) \right\|_2 \\ &\quad = \left\| \int_0^1 D^3 f(x_t)[\delta] \, dt \right\|_2 \\ &\quad \leq K \left\| \delta \right\|_2. \end{aligned}$$
This completes the proof. □

2 Proof of Theorem 5.1

The proof is based on the optimality conditions on $\hat{\beta}$ for the original problem, and those on $\check{\beta}$ for the restricted problem. We first observe that $\check{\beta}_n$ exists, since the function $x \mapsto \left\| x \right\|_1$ is coercive. Recall that $\check{\beta}_n$ is assumed to be uniquely defined.

To achieve sparsistency, it suffices that $\hat{\beta}_n = \check{\beta}_n$ and $\mathrm{supp}\, \check{\beta}_n = \mathrm{supp}\, \beta^*$. We derive sufficient conditions for $\hat{\beta}_n = \check{\beta}_n$ in Lemma 2.1, and make this sufficient condition explicitly dependent on the problem parameters in Lemma 2.2. This lemma will require that $\left\| \check{\beta}_n - \beta^* \right\|_2 \leq R_n$ for some $R_n > 0$. We will derive an estimation error bound of the form $\left\| \check{\beta}_n - \beta^* \right\|_2 \leq r_n$ in Lemma 2.4. We will then conclude that $\hat{\beta}_n = \check{\beta}_n$ if $r_n \leq R_n$ and the assumptions in Lemma 2.2 are satisfied, from which it will follow that $\mathrm{sign}\, \check{\beta} = \mathrm{sign}\, \beta^*$ provided that $\beta_{\min} \geq r_n$.

The following lemma is proved via an extension of the techniques of [Wainwright, 2009].

Lemma 2.1. *We have $\hat{\beta}_n = \check{\beta}_n$ if*
$$\left\| \left[\nabla L_n(\check{\beta}_n) \right]_{\mathcal{S}^c} \right\|_\infty < \tau_n. \qquad (3)$$

Proof. Recall that L_n is convex by assumption. Also recall that $\check{\beta}_n$ is assumed to be uniquely defined, and hence it is the only vector the satisfies the corresponding optimality condition:
$$\left[\nabla L_n(\check{\beta}_n) \right]_{\mathcal{S}} + \tau_n \check{z}_{\mathcal{S}} = 0 \qquad (4)$$

for some $\check{z}_{\mathcal{S}}$ such that $\|\check{z}_{\mathcal{S}}\|_\infty \leq 1$. Moreover, the fact that (3) is satisfied means that there exists $\check{z}_{\mathcal{S}^c}$ such that $\|\check{z}_{\mathcal{S}^c}\|_\infty < 1$ and

$$\nabla L_n(\check{\beta}_n) + \tau_n \check{z} = 0,$$

where $\check{z} := (\check{z}_{\mathcal{S}}, \check{z}_{\mathcal{S}^c})$. Therefore, $\check{\beta}_n$ is a minimizer of the original optimization problem in $\mathbb{R}^p$.

We now address the uniqueness of $\hat{\beta}$. By a similar argument to Lemma 1 in [Ravikumar et al., 2010] (see also Lemma 1(b) in [Wainwright, 2009]), any minimizer $\tilde{\beta}$ of the original optimization problem satisfies $\tilde{\beta}_{\mathcal{S}^c} = 0$. Thus, since $\check{\beta}$ is the only optimal vector for the restricted optimization problem, we conclude that $\hat{\beta}_n = \check{\beta}_n$ uniquely. □

We now combine Lemma 2.1 with the assumptions of Theorem 5.1 to obtain the following.

Lemma 2.2. *Under assumptions 1, 2, 3 and 6 of Theorem 5.1, we have $\hat{\beta}_n = \check{\beta}_n$ if $\check{\beta} \in \mathcal{N}_{\beta^*} \cap \mathcal{B}_{R_n}$, where $\mathcal{B}_{R_n} := \{\beta : \|\beta - \beta^*\|_2 \leq R_n, \beta_{\mathcal{S}^c} = 0, \beta \in \mathbb{R}^p\}$ with*

$$R_n = \frac{1}{2}\sqrt{\frac{\alpha \tau_n}{K}}. \tag{5}$$

Proof. Applying a Taylor expansion at β^*, and noting that both β^* and $\check{\beta}_n$ are supported on $\mathcal{S}$, we obtain

$$\begin{aligned}\left[\nabla L(\check{\beta}_n)\right]_{\mathcal{S}^c} &= \left[\nabla L_n(\beta^*)\right]_{\mathcal{S}^c} \\ &\quad + \left[\nabla^2 L_n(\beta^*)\right]_{\mathcal{S}^c,\mathcal{S}} \left(\check{\beta}_n - \beta^*\right)_{\mathcal{S}} \\ &\quad + (\epsilon_n)_{\mathcal{S}^c}\,, \end{aligned} \tag{6}$$

where the remainder term is given by

$$\epsilon_n = \int_0^1 (1-t) D^3 L_n(\beta_t)[\check{\beta} - \beta^*, \check{\beta} - \beta^*]\, dt$$

with $\beta_t := \beta^* + t(\check{\beta} - \beta^*)$ (see Section 4.5 of [Zeidler, 1995]), and thus satisfies

$$\|\epsilon_n\|_\infty \leq \sup_{t \in [0,1]} \left\{ \left\| D^3 L_n(\beta_t)[\check{\beta} - \beta^*, \check{\beta} - \beta^*] \right\|_\infty \right\}. \tag{7}$$

Recall the optimality condition for $\check{\beta}$ in (4). Again using a Taylor expansion, we can write this condition as

$$\begin{aligned}\left[\nabla L_n(\beta^*)\right]_{\mathcal{S}} + \left[\nabla^2 L_n(\beta^*)\right]_{\mathcal{S},\mathcal{S}} \left(\check{\beta}_n - \beta^*\right)_{\mathcal{S}} \\ + (\epsilon_n)_{\mathcal{S}} + \tau_n \check{z}_{\mathcal{S}} = 0. \end{aligned} \tag{8}$$

Recall that $\left[\nabla^2 L_n(\beta^*)\right]_{\mathcal{S},\mathcal{S}}$ is invertible by the second assumption of Theorem 5.1. Solving for $\left(\check{\beta}_n - \beta^*\right)_{\mathcal{S}}$ in (8) and substituting the solution into (6), we obtain

$$\begin{aligned}&\left[\nabla L_n(\check{\beta}_n)\right]_{\mathcal{S}^c} \\ &\quad = -\tau_n \left[\nabla^2 L_n(\beta^*)\right]_{\mathcal{S}^c,\mathcal{S}} \left[\nabla^2 L_n(\beta^*)\right]^{-1}_{\mathcal{S},\mathcal{S}} \check{z}_{\mathcal{S}} \\ &\qquad + \left[\nabla L(\beta^*)\right]_{\mathcal{S}^c} \\ &\qquad - \left[\nabla^2 L_n(\beta^*)\right]_{\mathcal{S}^c,\mathcal{S}} \left[\nabla^2 L_n(\beta^*)\right]^{-1}_{\mathcal{S},\mathcal{S}} \left[\nabla L_n(\beta^*)\right]_{\mathcal{S}} \\ &\qquad + (\epsilon_n)_{\mathcal{S}^c} \\ &\qquad - \left[\nabla^2 L_n(\beta^*)\right]_{\mathcal{S}^c,\mathcal{S}} \left[\nabla^2 L_n(\beta^*)\right]^{-1}_{\mathcal{S},\mathcal{S}} (\epsilon_n)_{\mathcal{S}}. \end{aligned}$$

Using the irrepresentability condition (assumption 3 of Theorem 5.1) and the triangle inequality, we have $\left\| \left[\nabla L_n(\check{\beta}_n)\right]_{\mathcal{S}^c} \right\|_\infty < \tau_n$ provided that

$$\max\left\{ \|\nabla L_n(\beta^*)\|_\infty, \|\epsilon_n\|_\infty \right\} \leq \frac{\alpha}{4}\tau_n.$$

The first requirement $\|\nabla L_n(\beta^*)\|_\infty \leq (\alpha/4)\tau_n$ is simply assumption 6 of Theorem 5.1, so it remains to determine a sufficient condition for $\|\epsilon_n\|_\infty \leq (\alpha/4)\tau_n$. Since L_n satisfies the $(\beta^*, \mathcal{N}_{\beta^*})$-LSSC with parameter K, we have from (7) that

$$\|\epsilon_n\|_\infty \leq K \left\| \check{\beta} - \beta^* \right\|_2^2,$$

provided that $\check{\beta} \in \mathcal{N}_{\beta^*}$ (since $\mathcal{N}_{\beta^*}$ is convex by assumption, this implies $\beta_t \in \mathcal{N}_{\beta^*}$). Thus, to have $\|\epsilon_n\|_\infty \leq \frac{\alpha}{4}\tau_n$, it suffices that

$$\left\| \check{\beta} - \beta^* \right\|_2 \leq \frac{1}{2}\sqrt{\frac{\alpha \tau_n}{K}}$$

and $\check{\beta} \in \mathcal{N}_{\beta^*}$. □

To bound the distance $\left\| \check{\beta} - \beta^* \right\|_2$, we adopt an approach from [Ravikumar et al., 2010, Rothman et al., 2008]. We begin with an auxiliary lemma.

Lemma 2.3. *Let $g : \mathbb{R}^p \to \mathbb{R}$ be a convex function, and let $z \in \mathbb{R}^p$ be such that $g(z) \leq 0$. Let $\mathcal{B} \subset \mathbb{R}^p$ be a closed set, and let $\partial\mathcal{B}$ be its boundary. If $g > 0$ on $\partial\mathcal{B}$ and $g(b) \leq 0$ for some $b \in \mathcal{B} \setminus \partial\mathcal{B}$, then $x \in \mathcal{B}$.*

Proof. We use a proof by contradiction. Suppose that $z \notin \mathcal{B}$. We first note that there exists some $t^* \in (0,1)$ such that $b + t^*(z - b) \in \partial\mathcal{B}$; if such a t^* did not exist, then we would have $z_t := b + t(z - b) \to z$ as $t \to 1$, which is impossible since $z \notin \mathcal{B}$ and $\mathcal{B}$ is closed.

We now use the convexity of g to write

$$g(b + t^*(x - b)) \leq (1 - t^*)g(b) + t^* g(x) \leq 0,$$

which is a contradiction since $g > 0$ on $\partial\mathcal{B}$. □

The following lemma presents the desired bound on $\left\| \check{\beta}_n - \beta^* \right\|_2$; note that this can be interpreted as the estimation error in the $n > p$ setting, considering $\beta^*_{\mathcal{S}}$ as the parameter to be estimated.

Lemma 2.4. *Define the set*

$$\mathcal{B}_{r_n} := \left\{ \beta \in \mathbb{R}^p : \left\| \beta - \beta^* \right\|_2 \leq r_n, \beta_{\mathcal{S}^c} = 0 \right\},$$

where

$$r_n := \frac{\alpha + 4}{\lambda_{\min}} \sqrt{s} \tau_n. \tag{9}$$

Under assumptions 1, 2, 6 and 7 of Theorem 5.1, if

$$\tau_n < \frac{3\lambda_{\min}^2}{2(\alpha+4)Ks}, \tag{10}$$

then $\check{\beta}_n \in \mathcal{B}_{r_n}$.

Proof. Set $s = |\mathcal{S}|$, and for $\beta \in \mathbb{R}^s$ let $Z(\beta) = (\beta, 0) \in \mathbb{R}^p$ be the zero-padding mapping, where $(\beta, 0)$ denotes the vector that equals to β on $\mathcal{S}$ and 0 on $\mathcal{S}^{\mathrm{c}}$. Then we have

$$\check{\beta}_{\mathcal{S}} = \arg\min_{\beta \in \mathbb{R}^s} \left\{ (L_n \circ Z)(\beta) + \tau_n \left\| \beta \right\|_1 \right\}.$$

For $\delta \in \mathbb{R}^s$, define

$$g(\delta) = (L_n \circ Z)(\beta^*_{\mathcal{S}} + \delta) - (L_n \circ Z)(\beta^*_{\mathcal{S}}) + \tau_n \left(\|\beta^*_{\mathcal{S}} + \delta\|_1 - \|\beta^*_{\mathcal{S}}\|_1 \right).$$

We trivially have $g(0) = 0$, and thus $g(\delta^*) \leq g(0) = 0$, where $\delta^* := \check{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}$. Now our goal is prove that $g > 0$ on the boundary of $(\mathcal{B}_{r_n})_{\mathcal{S}} := \{ \delta \in \mathbb{R}^s : \|\delta\|_2 \leq r_n \}$, thus permitting the application of Lemma 2.3.

We proceed by deriving a lower bound on $g(\delta)$. We define $\phi(t) := (L_n \circ Z)(\beta^*_{\mathcal{S}} + t\delta)$, and write the following Taylor expansion:

$$\begin{aligned} &(L_n \circ Z)(\beta^*_{\mathcal{S}} + \delta) - (L_n \circ Z)(\beta^*_{\mathcal{S}}) \\ &\quad = \phi(1) - \phi(0) \\ &\quad = \phi'(0) + \frac{1}{2}\phi''(0) + \frac{1}{6}\phi'''(\tilde{t}), \end{aligned}$$

for some $\tilde{t} \in [0, 1]$ (recall that L_n is three times differentiable by assumption). We bound the term $\phi'(0)$ as follows:

$$\begin{aligned} |\phi'(0)| &= \left| \left\langle \left[\nabla L_n(\beta^*) \right]_{\mathcal{S}}, \delta \right\rangle \right| \\ &\leq \sqrt{s} \left\| \left[\nabla L_n(\beta^*) \right]_{\mathcal{S}} \right\|_\infty \|\delta\|_2 \\ &\leq \frac{\alpha \tau_n}{4} \sqrt{s} \|\delta\|_2, \end{aligned}$$

where the first step is by Hölder's inequality and the identity $\|z\|_2 \leq \sqrt{s}\|z\|_1$, and the second step uses assumption 6 of Theorem 5.1. To bound the term $\phi''(0)$, we use the second assumption of Theorem 5.1 to write

$$\phi''(0) = \delta^T \left[\nabla^2 L_n(\beta^*) \right]_{\mathcal{S},\mathcal{S}} \delta \geq \lambda_{\min} \|\delta\|_2^2.$$

We now turn to the term $\phi'''(\tilde{t})$. Again using the fact that L_n satisfies the $(\beta^*, \mathcal{N}_{\beta^*})$-LSSC with parameter K, it immediately follows that $(L_n \circ Z)$ satisfies the $(\beta^*_{\mathcal{S}}, (\mathcal{N}_{\beta^*})_{\mathcal{S}})$-LSSC with parameter K, where $(\mathcal{N}_\beta)_{\mathcal{S}} = \{ \beta_{\mathcal{S}} : \beta \in \mathcal{N}_{\beta^*} \}$. Hence, and also making use of Hölder's inequality and the fact that $\|z\|_1 \leq \sqrt{s}\|z\|_2$ ($z \in \mathbb{R}^s$), we have

$$\begin{aligned} \left| \phi'''(\tilde{t}) \right| &= \left| D^3 (L_n \circ Z)(\beta^*_{\mathcal{S}} + \tilde{t}\delta)[\delta, \delta, \delta] \right| \\ &\leq \|\delta\|_1 \left\| D^3 (L_n \circ Z)(\beta^*_{\mathcal{S}} + \tilde{t}\delta)[\delta, \delta] \right\|_\infty \\ &\leq K\sqrt{s} \|\delta\|_2^3 \end{aligned}$$

provided that $\beta^*_{\mathcal{S}} + \tilde{t}\delta \in (\mathcal{N}_\beta)_{\mathcal{S}}$. Since $\mathcal{B}_{r_n} \subseteq \mathcal{N}_{\beta^*}$ by assumption 7 of Theorem 5.1, the latter condition holds provided that $\delta \in (\mathcal{B}_{r_n})_{\mathcal{S}}$.

Using the triangle inequality, we have

$$\left| \|\beta^*_{\mathcal{S}} + \delta\|_1 - \|\beta^*_{\mathcal{S}}\|_1 \right| \leq \|\delta\|_1 \leq \sqrt{s} \|\delta\|_2.$$

Hence, and combining the preceding bounds, we have $g(\delta) \geq f(\|\delta\|_2)$, where

$$f(x) = -\frac{\alpha \tau_n}{4} \sqrt{s} x + \frac{\lambda_{\min}}{2} x^2 - \frac{K\sqrt{s}}{6} x^3 - \sqrt{s} \tau_n x.$$

Observe that if the inequality

$$0 < x < \frac{3\lambda_{\min}}{2K\sqrt{s}}. \tag{11}$$

holds, then we can bound the coefficient to x^3 in terms of that of x^2 to obtain

$$f(x) > \frac{\lambda_{\min}}{4} x^2 - \left(1 + \frac{\alpha}{4} \right) \sqrt{s} \tau_n x. \tag{12}$$

By a direct calculation, this lower bound has roots at 0 and r_n (see (9)), and hence $f(r_n) > 0$ provided that $x = r_n$ satisfies (11). By a direct substitution, this condition can be ensured by requiring that

$$\tau_n < \frac{3\lambda_{\min}^2}{2(\alpha+4)Ks}. \tag{13}$$

Recalling that $g(\delta) \geq f(\|\delta\|_2)$, we have proved that g satisfies the conditions of Lemma 2.3 with $z = \delta^*$, $b = 0$, and $\mathcal{B} = (\mathcal{B}_{r_n})_{\mathcal{S}}$, and we thus have $\delta^* \in (\mathcal{B}_{r_n})_{\mathcal{S}}$, or equivalently $\check{\beta}_n \in \mathcal{B}_{r_n}$.

□

We now combine the preceding lemmas to obtain Theorem 5.1. We require $r_n \leq R_n$ so the assumption that $\left\| \check{\beta} - \beta^* \right\|_\infty \leq R_n$ in Lemma 2.2 is satisfied. From the definitions in (5) and (9), this is equivalent to requiring

$$\tau_n \leq \frac{\lambda_{\min}^2}{4(\alpha+4)^2} \frac{\alpha}{Ks},$$

which is true by assumption 5 of the theorem. This assumption also implies that (10) holds, since $\frac{\alpha}{4(\alpha+4)} \leq \frac{3}{2}$ for any $\alpha \geq 0$. Finally, by the conclusion of Lemma 2.4, we have successful sign pattern recovery if $\beta_{\min} \geq r_n$, thus recovering assumption 4 of the theorem.

3 Proofs of the Results in Section 6

3.1 Proof of Corollary 6.2

By a direct differentiation, we obtain for $j \in \{1,\dots,p\}$ that

$$[\nabla L_n(\beta^*)]_j = -\sum_{i=1}^{n} \varepsilon_i (x_i)_j,$$

where $\varepsilon_i = n^{-1}\left(Y_i - \mathsf{E}\, Y_i\right)$.

Fix $j \in \{1,\dots,p\}$, and let $X_i := n^{-1}(x_i)_j Y_i$. As $X_1,\dots,X_n$ are bounded, they can be characterized using Hoeffding's inequality [Boucheron et al., 2013].

Theorem 3.1 (Hoeffding's Inequality)**.** *Let $X_1,\dots,X_n$ be independent random variables such that X_i takes its value in $[a_i,b_i]$ almost surely for all $i \in \{1,\dots,n\}$. Then*

$$\begin{aligned}&\mathsf{P}\left\{\left|\sum_{i=1}^{n}(X_i - \mathsf{E}\, X_i)\right| \geq t\right\}\\ &\quad\leq 2\exp\left[-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right].\end{aligned}$$

In our case, we can set $(b_i-a_i)^2 = n^{-2}(x_i)_j^2$, since $Y_i \in \{0,1\}$. Since $\sum_{i=1}^{n}|(x_i)_j|^2 \leq n$ for all k by assumption, we obtain

$$\sum_{i=1}^{n}(b_i-a_i)^2 \leq \frac{1}{n}. \tag{14}$$

Thus, by Hoeffding's inequality and the union bound, we obtain

$$\begin{aligned}&\mathsf{P}\left\{\|\nabla L_n(\beta^*)\|_\infty \geq \frac{\alpha\tau_n}{4}\right\}\\ &\quad\leq \sum_{j=1}^{p}\mathsf{P}\left\{\left|[\nabla L_n(\beta^*)]_j\right| \geq \frac{\alpha\tau_n}{4}\right\}\\ &\quad\leq 2\exp\left(\ln p - 2nt^2\right)\Big|_{t=\frac{\alpha\tau_n}{4}}.\end{aligned}$$

This decays to zero provided that $\tau_n \gg (n^{-1}\log p)^{1/2}$. Substituting this scaling into the fifth condition of Theorem 5.1, we obtain the condition $s^2(\log p)\,\nu_n^4\gamma_n^2 \ll n$. The required uniqueness of $\check{\beta}$ can be proved by showing that the composition $L_n \circ Z$ (with Z being the zero-padding of a vector in $\mathbb{R}^s$) is strictly convex, given the second condition of Theorem 5.1. One way to prove this is via self-concordant like inequalities [Tran-Dinh et al., 2013]; we omit the proof here for brevity.

3.2 Proof of Corollary 6.3

Let $Y_1,\dots,Y_n$ be independent gamma random variables with shape parameter $k>0$ and scale parameter θ_i respectively. We have, for $q \in \mathbb{N}$,

$$\mathsf{E}\,|Y_i|^q = \frac{\Gamma(q+k)}{\Gamma(k)}\theta_i^q,$$

where Γ denotes the gamma function.

To study the concentration of measure behavior of $\nabla L_n(\beta^*)$, we use the following result [Boucheron et al., 2013].

Theorem 3.2 (Bernstein's Inequality)**.** *Let $X_1,\dots,X_n$ be independent real random variables. Suppose that there exist $v>0$ and $c>0$ such that $\sum_{i=1}^{n}\mathsf{E}\,X_i^2 \leq v$, and*

$$\sum_{i=1}^{n}\mathsf{E}\,|X_i|^q \leq \frac{q!}{2}vc^{q-2}$$

for all integers $q \geq 3$. Then

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

We proceed by evaluating the required moments for our setting. By a direct differentiation, we obtain

$$[\nabla L_n(\beta^*)]_j = \sum_{i=1}^{n}\varepsilon_i\,(x_i)_j$$

for $j \in \{1,\dots,p\}$, where $\varepsilon_i := n^{-1}\left(Y_i - \mathsf{E}\,Y_i\right)$.

Fix $j \in \{1,\dots,p\}$, and let $X_i := n^{-1}(x_i)_j Y_i$. We have

$$\begin{aligned}\sum_{i=1}^{n}\mathsf{E}\,X_i^2 &= \sum_{i=1}^{n}\frac{(x_i)_j^2}{n^2}\mathsf{E}\,Y_i^2\\ &= \sum_{i=1}^{n}\frac{(x_i)_j^2}{n^2}\frac{\Gamma(k+2)}{\Gamma(k)}\theta_i^2.\end{aligned}$$

Recall that $\theta_i = k^{-1}\langle x_i,\beta^*\rangle^{-1}$. Using the first displayed equation in Section 7.3, we have

$$\theta_i \leq (k\mu_n)^{-1}, \tag{15}$$

and thus

$$\begin{aligned}\sum_{i=1}^{n}\mathsf{E}\,X_i^2 &\leq \frac{1}{(n\mu_n)^2}\frac{\Gamma(k+2)}{k^2\Gamma(k)}\sum_{i=1}^{n}\frac{(x_i)_j^2}{\|x_i\|_2^2}\\ &\leq \frac{1}{n\mu_n^2}\frac{\Gamma(k+2)}{k^2\Gamma(k)},\end{aligned}$$

where we have applied the assumption $\sum_{i=1}^{n}(x_i)_j^2 \leq n$. Using the identity $\Gamma(k+2) = k(k+1)\Gamma(k)$, we obtain

$$\sum_{i=1}^{n}\mathsf{E}\,X_i^2 \leq \frac{k+1}{n\mu_n^2 k}.$$

As for the moments of higher orders, we have

$$\begin{aligned}\sum_{i=1}^{n} \mathsf{E}\,|X_i|^q &= \sum_{i=1}^{n} \frac{|(x_i)_j|^q}{n^q} \mathsf{E}\,|Y_i|^q \\ &= \sum_{i=1}^{n} \frac{|(x_i)_j|^q}{n^q} \frac{\Gamma(k+q)}{\Gamma(k)} \theta_i^q.\end{aligned}$$

With the upper bound (15) on θ_i, we have

$$\begin{aligned}\sum_{i=1}^{n} \mathsf{E}\,|X_i|^q &\leq \frac{\Gamma(k+q)}{(kn\mu_n)^q \Gamma(k)} \sum_{i=1}^{n} |(x_i)_j|^q \\ &= \frac{\Gamma(k+q)}{(kn\mu_n)^q \Gamma(k)} \left\| ((x_1)_j, \ldots, (x_n)_j) \right\|_q^q.\end{aligned}$$

Using the identity $\|z\|_q \leq \|z\|_2$ for $q \geq 2$, and the assumption $\sum_{i=1}^{n} (x_i)_j^2 \leq n$, we obtain

$$\sum_{i=1}^{n} \mathsf{E}\,|X_i|^q \leq \frac{\Gamma(k+q)}{(k\sqrt{n}\mu_n)^q \Gamma(k)}.$$

For $k \in (0,1]$, we have $\frac{\Gamma(k+q)}{\Gamma(q)} \leq q!$, and hence by a direct substitution it suffices to choose

$$v = \frac{k+1}{n\mu_n^2 k^2}, \quad c = \frac{1}{k\sqrt{n}\mu_n}. \tag{16}$$

For $k \in (1,\infty)$, we have by induction on q that $\frac{\Gamma(k+q)}{\Gamma(q)} \leq q! k^q$. Thus, for $k \in (1,\infty)$, it suffices that

$$v = \frac{2k}{n\mu_n^2}, \quad c = \frac{1}{\sqrt{n}\mu_n}. \tag{17}$$

Thus, applying Bernstein's inequality and the union bound, we obtain

$$\begin{aligned}&\mathsf{P}\left\{ \|\nabla L_n(\beta^*)\|_\infty \geq \frac{\alpha\tau_n}{4} \right\} \\ &\leq \sum_{i=1}^{p} \mathsf{P}\left\{ |[\nabla L_n(\beta^*)]_i| \geq \frac{\alpha\tau_n}{4} \right\} \\ &\leq 2\exp\left[\ln p - \frac{t^2}{2(v+ct)} \right] \Bigg|_{t=\frac{\alpha\tau_n}{4}}.\end{aligned}$$

Since L_n is self-concordant and $\left[D^2 L_n(\beta^*) \right]_{\mathcal{S},\mathcal{S}}$ is positive definite by assumption, the composition $L_n \circ Z$ with the padding operator Z is strictly convex [Nesterov, 2004, Nesterov and Nemirovskii, 1994] and thus $\check{\beta}_n$ uniquely exists. Therefore, we can apply Theorem 5.1. The scaling laws on τ_n and (p,n,s) follow via the same argument to that in the proof of Corollary 6.2. Note that the final condition of Theorem 5.1 also imposes conditions on (p,n,s), but for this term even the weaker condition $s^2(\log p)\nu_n^2 \ll n$ suffices.

4 Proof of Corollary 6.4

By a direct differentiation, we obtain

$$\nabla L_n(\Theta^*) = \hat{\Sigma}_n - (\Theta^*)^{-1} = \hat{\Sigma}_n - \Sigma.$$

We apply the following lemma from [Ravikumar et al., 2011] to study the concentration behavior of $\nabla L_n(\Theta^*)$.

Lemma 4.1. *Let Σ and $\hat{\Sigma}_n$ be defined as in Section 6.4. We have*

$$\begin{aligned}&\mathsf{P}\left\{ \left| \left(\hat{\Sigma}_n \right)_{i,j} - \Sigma_{i,j} \right| > t \right\} \\ &\leq 4\exp\left[-\frac{nt^2}{128(1+4c^2)^2 \kappa_{\Sigma^*}^2} \right],\end{aligned}$$

for all $t \in (0, 8\kappa_{\Sigma^*}(1+c)^2)$.

Using the union bound, we have

$$\begin{aligned}&\mathsf{P}\left\{ \|\nabla L_n(\Theta^*)\|_\infty \leq \frac{\alpha\tau_n}{4} \right\} \\ &\leq 4p^2 \exp\left[-\frac{nt^2}{128(1+4\sigma^2)^2 \kappa_{\Sigma^*}^2} \right] \Bigg|_{t=\frac{\alpha\tau_n}{4}},\end{aligned}$$

provided that $\tau_n \to 0$, and that n is large enough so that the upper bound on t in the lemma is satisfied.

Define

$$\begin{aligned}\check{\Theta}_n \in \arg\min_{\Theta} \{ &L_n(\Theta) + \tau_n |\Theta|_1 : \\ &\Theta > 0, \Theta_{\mathcal{S}^c} = 0, \Theta \in \mathrm{R}^{p\times p} \}.\end{aligned} \tag{18}$$

Since L_n is self-concordant and $\left[D^2 L_n(\Theta^*) \right]_{\mathcal{S},\mathcal{S}}$ is positive definite by assumption, the composition $L_n \circ Z$ with the padding operator Z is strictly convex [Nesterov, 2004, Nesterov and Nemirovskii, 1994] and thus $\check{\Theta}_n$ uniquely exists. Therefore, we can apply Theorem 5.1. The scaling laws on τ_n and (p,n,s) follow via the same arguments as the preceding examples.

References

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford Univ. Press, Oxford.

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

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

[Ravikumar et al., 2010] Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Stat.*, 38(3):1287–1319.

[Ravikumar et al., 2011] Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980.

[Rothman et al., 2008] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Elect. J. Stat.*, 2:494–515.

[Tran-Dinh et al., 2013] Tran-Dinh, Q., Li, Y.-H., and Cevher, V. (2013). Minimization of self-concordant like functions.

[Wainwright, 2009] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory*, 55(5):2183–2202.

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