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
$$\begin{aligned} \left| D^3 f(x^* + \delta)[u, u, u] \right| &= \left| \langle u, H u \rangle \right| \\ &\leq \left\| H \right\|_2 \left\| u \right\|^2, \end{aligned}$$
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 $\|\check{z}_{\mathcal{S}}\|_\infty \le 1$. 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 $\mathcal{B}_{R_n} := \{\beta : \|\beta - \beta^*\|_2 \le R_n, \beta_{\mathcal{S}^c} = 0, \beta \in \mathbb{R}^p\}$ 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

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

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

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

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

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

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

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

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

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

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

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

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 $\|\epsilon_n\|_\infty \le \frac{\alpha}{4}\tau_n$, it suffices that

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

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 $g(z) \le 0$. 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 $g(b) \le 0$ 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

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

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 $\mathcal{S}^c$. 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:

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

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 $g(\delta) \geq f\left(\|\delta\|_2\right)$, 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 $g(\delta) \geq f\left(\|\delta\|_2\right)$, 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 $j \in \{1, \ldots, p\}$ that

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

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

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

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

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

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

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

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 $Y_1, \ldots, Y_n$ 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 $X_1, \ldots, X_n$ 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*

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

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

Fix $j \in \{1, \ldots, p\}$, 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

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

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

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

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.